RESEARCH REPORT

Serial No. 10

May 1992

MODELLING THE WINNING PROBABILITIES

by

J. Bacon-Shone, V.S.Y. Lo and K. Busche

THE UNIVERSITY OF HONG KONG

DEPARTMENT OF STATISTICS

Pokfulam Road, Hong Kong. Tel: (852) 859 2466 Fax: (852) 858 9041

E-Mail: STATIST@HKUCC.HKU.HK or STATIST@HKUCC.BITNET

MODELLING THE WINNING PROBABILITIES

John Bacon-Shone — *Dept. of Statistics*
Victor S.Y. Lo — *Dept. of Statistics*
Kelly Busche — *School of Economics*

University of Hong Kong

Abstract

Previous studies conclude that a favourite-longshot bias exists in win betting at horse race tracks and this is interpreted as risk-preferring behaviour. However, the statistical techniques used contain certain weaknesses, including the neglect of the necessary correlation between bet fractions which must sum to one in every race. We propose some classes of logit models to analyse the relationship between winning probabilities and the bet fractions. A simple logit model is proposed after going through a modelling process and, by Cox's test, it is preferred to previous models (Ali (1977) and Asch, Malkiel & Quandt (1984)). Empirical results are obtained for several racetracks in the U.S., Hong Kong, Japan, and China. No strong conclusion for risk preference can be maintained.

Acknowledgement

We thank Professors M.M. Ali and Richard Quandt for their data, and Professor Junji Shiba for his help in acquiring the Japanese data.

I. Introduction

Research into gambling at racetracks has not concentrated on refined modelling of individual betting behaviour. Rather, relatively simple models are used to measure efficiency of the racetrack market and risk preference of gamblers. In studying these issues, researchers (including Ali (1977), Synder (1978), Asch, Malkiel and Quandt (1982), Hausch Ziemba and Rubinstein (1981), and Busche and Hall (1988)) have focused on bet fractions as estimators of the winning probabilities. Based on rudimentary statistical analysis, almost all (the sole exception being Busche and Hall) concluded that there was evidence for a favourite-longshot bias (sometimes referred to as risk preference), i.e.

bettors underbet favourites and overbet longshots relative to equal marginal return bets. Thaler & Ziemba (1988, p.163) refer to the favourite-longshot bias as "the most robust anomalous empirical regularity". The method chosen by all these researchers has certain weaknesses : first, there may be a classification problem related to the grouping of the data by the dependent variable (see Busche and Hall (1988)); and second, no account has been taken of the unit sum constraint on bet fractions, i.e., since bet fractions sum to one, they cannot be independent but are always treated as such. In this paper, we suggest a different statistical technique to analyse the win bet data, and apply that technique to previously analyzed data from the U.S. and Hong Kong and new data from Japan and Shanghai. With a model which is independent of classification errors and explicitly takes the unit sum into account, we find that no strong conclusion for a longshot bias or risk preference can be maintained.

In Part II, we review the traditional model and the empirical findings, and in Part III, we provide an alternative class of models where classification and unit constraint problems are explicitly taken into account. In Part IV we test each model in the class using the already analyzed data of Ali(1977) and choose one preferred model. In Part V, we fit the preferred model on several other data sets, some of which have been analyzed elsewhere and some of which are new. Comparisons with other models are given in Part VI. Conclusions are in Part VII.

II. The Traditional method of analysing win bet data

To study issues of risk preference and market efficiency, researchers (we use Ali(1977) as the archetype) have assumed individual utility maximizers each betting a small fixed amount on the outcome of a race. Betting is treated as a means to an end, not an end in itself, and betting is parimutuel, i.e. the odds are determined by relative amounts bet on all horses. In this setting, and if bettors anticipate final odds in the market, final odds will correctly reflect marginal utilities. Assuming equal marginal utilities, Ali(1977) estimated a utility of wealth function.

For each horse race, Ali (1977) classifies the horses by favourite position (favouritism). That is,

$$P_{(1)r} \geq \dots \geq P_{(n_r)r}$$

where P_{ir} is the win bet fraction of horse i in race r, and n_r = total number of horses in race r.

Average bet fraction for the ith favourite = $\overline{P}_{(i)} = \frac{1}{m} \sum_{r=1}^{m} P_{(i)r}$

where m = total number of races in the data set.

Further, we define the following indicator variable :

$$Y_{ir} = \begin{cases} 1 & \text{if horse i wins race r} \\ 0 & \text{otherwise} \end{cases}$$

Obviously, there is only one '1' in $(Y_{1r},\dots,Y_{n_r r})$.

Also, $Y_{(i)r}$ is the indicator variable associated with $P_{(i)r}$.

The observed winning frequency for the ith favourite is :

$$\overline{Y}_{(i)} = \frac{1}{m} \sum_{r=1}^{m} Y_{(i)r} \qquad \forall\ i$$

To test whether the bet fractions are good estimates of the winning probabilities, Ali compared the bet fractions and the observed winning frequencies by simple Z-statistics :

$$Z_i = \frac{\overline{P}_{(i)} - \overline{Y}_{(i)}}{\hat{se}\ (\overline{Y}_{(i)})} \qquad \forall\ i$$

where $\hat{se}\ (\overline{Y}_{(i)}) = \sqrt{\overline{Y}_{(i)}(1-\overline{Y}_{(i)})\ /\ m}$

Ali's results are shown in table 1.

Table 1

Illustration of Ali's method using Ali's data

fav.pos. i	#races	bet frac $\overline{P}_{(i)}$	winfreq $\overline{Y}_{(i)}$	$\hat{s.e.}$	Z_i
1	20247	0.3245	0.3513	0.0034	-7.9868
2	20247	0.2078	0.2060	0.0028	0.6648
3	20247	0.1514	0.1557	0.0025	-1.6735
4	20247	0.1122	0.1056	0.0022	3.0503
5	20232	0.0828	0.0773	0.0019	2.9462
6	20088	0.0602	0.0554	0.0016	2.9462
7	19281	0.0418	0.0347	0.0013	5.3414
8	15749	0.0276	0.0206	0.0011	6.1862
9	299	0.0158	0.0033	0.0033	3.7388
10	71	0.0117	0.0141	0.0140	-0.1683

(Note : the result above is trivially different from that reported in Ali (1977). Where there were equal odds between the winning horse and some other horses, he always chose the winning horse as a more favourite horse, thus biasing the results slightly towards the overbetting of favourites result. We use an alternative method : we alternate between the first and the second horses as the more favourite horse whenever there is a tie.)

The Z_i statistics suggest the existence of a favourite-longshot bias, i.e. the bettors underbet favourites and overbet longshots.

Some weaknesses of Ali's method are suggested :

1. Bet fractions in Ali's analysis are treated as independent across horses. The constraint that bet fractions must sum to one introduces a necessary but neglected correlation between bet fractions.

2. Ali's method classifies data by the dependent variable (bet fraction) with unknown results (See Busche and Hall (1988) for discussion). Moreover, the 6th favourite in a 6-horse-race may be quite different from

a 6th favourite in a 10-horse-race.

3. When using the Z-test, Ali treated the average bet fraction as a constant value rather than a random variable. A random variable may be more appropriate here since $\overline{P}_{(i)}$ is essentially an estimator of the probability of the ith favourite horse.

4. Ali used the Binomial distribution for the winning event of a horse. To utilize the information more fully, the Multinomial distribution would seem more appropriate. (This has been noted by Asch and Quandt (1984) in their model.)

III. Classes of Multinomial Logit Models

We consider two classes of multinomial logit models to fit the winning probability on the bet fraction. (See Li (1986) or Chapter 12 of Aitchison (1986)).

We start with a model of the following form (omitting the subscript r for race r) :

$$\ln(\pi_{(i)}/\pi_{(n)}) = \alpha_i + \beta_i \ln(P_{(i)}/P_{(n)}) \qquad i = 1,2,\ldots,n-1 \qquad (1)$$

where. $P_{(i)}$ = bet fraction of the ith favourite and

$\pi_{(i)}$ = winning probability of the ith favourite.

We call this the A-class since the logit transform, $\ln(P_{(i)}/ P_{(n)})$, is commonly used by Aitchison (1986). In this class of model, the logit (in the multivariate case) of the winning probabilities is linearly related to the logit of the bet fractions. Note however, that the results are dependent on the arbitrary choice of the divisor (horse n in (1) above) since all measures are relative to that horse.

Another similar class is :

$$\pi_{(i)} = \frac{\exp(\alpha_i + \beta_i \ln P_{(i)})}{\sum_j \exp(\alpha_j + \beta_j \ln P_{(j)})} \qquad (2)$$

(with $\alpha_n = 0$) $\qquad i=1,2,\ldots,n$

We call this the L-class. It is one of the common forms of the Polytomous model (see, for example, Hosmer and Lemeshow (1989)) or the Multinomial logit model (see, for example, Judge, Griffiths, Hill, Lutkepohl & Lee (1985) or Aitchison (1986)).

Both classes are multivariate extensions of the binary logit model, but in general, the A-class depends on which horse is chosen as the divisor. Thus the L-class model has the advantage of symmetry of parameters since no divisor is required. For estimation, we assume :

$$\underline{Y}_r = (Y_{1r},\ldots,Y_{n_r r}) \sim \text{Multinomial } (\underline{\pi}_r).$$

The likelihood function is :

$$\text{lik} = \prod_{r=1}^{m} \prod_{i=1}^{n_r} \pi_{ir}^{y_{ir}}$$

and the log likelihood can be written as :

$$l = \sum_{r=1}^{m} \ln \pi_{[1]r} \qquad (3)$$

where $\pi_{[1]r}$ denotes the winning probability associated the winning horse in race r. Maximum Likelihood Estimation can be easily applied to maximize (3) with respect to the parameters.

Notice that in both classes, if $\alpha_i = 0$ and $\beta_i = 1$, then $\pi_i = P_i$. This implies winning probabilities equal bet fractions, i.e. risk neutrality (See Ali(1977)).

Quadratic models are also tried, but the additional parameter is found to be significant in 8-horse-races only. The quadratic models fitted are :

A-class :

$$\ln(\pi_{(i)}/\pi_{(n)}) = \alpha_i + \beta_i \ln(P_{(i)}/P_{(n)}) + \nu\, [\ln(P_{(i)}/P_{(n)})]^2$$

$$i=1,2,\ldots,n-1$$

L-class :

$$\pi_{(i)} = \frac{\exp[\ \alpha_i + \beta_i \ \ln P_{(i)} + \nu\ (\ln P_{(i)})^2\ ]}{\sum_j \exp[\ \alpha_j + \beta_j \ \ln P_{(j)} + \nu\ (\ln P_{(j)})^2\ ]}$$

$$i = 1, 2, \ldots, n$$

(with $\alpha_n = 0$)

IV. Selection of models

Both the A-class and the L-class models are examined using some large data sets. We ignore the 5,9 and 10-horse races in Ali's data for the purpose of selection of models because they each have fewer than 300 races. The total number of races for 6, 7 and 8-horse races are 807, 3532 and 15450 respectively.

The results including quadratic models are shown in Figures 1-3 and 4-6 for A and L class models respectively. In these figures, models are arranged according to complexity. The top box is the simplest model and the lowest box is the most complicated model. E.g. 0, 1 inside the top box in each figure means $\alpha = 0$ and $\beta = 1\ \forall$ i. Models with constant α in the L-class are ignored since this implies setting one horse apart. The negative numbers inside the boxes are the associated log likelihoods and thus likelihood ratio tests can be used. The numbers on the lines are the associated p-values for the tests that the next models are better than the previous ones. AIC (Akaike Information Criteria) values are given in parentheses below the log likelihoods.

Within these classes of models, from the properties of the models and the results obtained, we prefer what we call the constant-β model. That is, the constant-β model has $\alpha_i = 0$ and $\beta_i = \beta\ \ \forall$ i. In this circumstance, both A and L classes reduce to the same constant-β model. The reasons we prefer this are :

1. The model is independent of the classification.

2. The sum-to-one constraint on the objective probabilities is explicitly taken into account.

3. β's are significantly different from one at 5% significance level in

all cases here. Also, the reduction of log likelihood caused by constant-β in the model fitting process is great in general. Moreover, using the Akaike Information Criterion (AIC), (i.e. taking the number of estimated parameters into account, see e.g. Sakamoto, Ishiguro & Kitagawa (1986)) the constant-β model attains the minimum values among all the fitted models on both 6-horse-races and 7-horse-races in both A-class and L-class. With 8-horse-races, models with more parameters achieve slightly better AIC values.

4. In addition, the estimated values of β are quite consistent in races with different race-sizes as is the case when Ali's data is broken into race sizes and analysed by his method. (The $\hat{\beta}$'s in the 6, 7 and 8-horse-races are 1.1426, 1.1656 and 1.1351 respectively.)

The constant-β model can be rearranged as :

$$Pr(Y_i = 1 \mid \underline{P}) = \pi_i = \frac{P_i^{\beta}}{\sum_j P_j^{\beta}} \quad \text{for } i=1,2,\ldots,n \tag{4}$$

Note that there is no requirement of any ordering convention.

β can be thought of measuring the "Bias" of the subjective probability with respect to the objective probability. We would say that the subjective probabilities are unbiased if β is not significantly different from one, i.e. $\pi_i = P_i$ which is the risk neutral hypothesis tested by almost all researchers. If β is less than one, then it can be interpreted as risk aversion. If β is greater than one, we can say that the bettors underbet favourites and overbet longshots. By using this model for analysis, the weaknesses of Ali's method disappear.

V. Empirical analyses in US, HK, Japan and Shanghai

We have fitted the constant-β model in different racetracks in the U.S.. The estimated β ranges from 1.102 to 1.156 in all racetracks and almost all of them are significantly different from one at a 5% level of significance. In Japan, the β is also significantly greater than one. However, in Hong Kong and Shanghai data sets, the β's are not significantly greater than one at any reasonable level (a result consistent with Busche & Hall (1988)). Overall, the estimated β's are close to one. The empirical results are summarized in table 2.

Table 2

Summary of β-model for win bet in different racetracks

Racetrack	Country	#races	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	p-value	Ave.win pool size (US$)
Happy Valley (81-89)	HK	2212	1.0372	.03256	.077	1,083,653
Shatin (81-89)	HK	1943	0.9359	.03060	.982	1,088,536
Shatin + Happy Valley (81-89)	HK	4155	0.9844	.02232	.758	1,085,936
Ali's (70-74)						
Saratoga	NY,US	9072	1.1562	.01642	≃ 0	24,790*
Roosevlt	NY,US	5806	1.1261	.02076	≃ 0	218,128*
Yonkers	NY,US	5369	1.1313	.02116	≃ 0	228,155*
Quandt's (83-84)						
Atlantic City	US	712	1.1023	.05912	.042	unknown
Meadowlands	US	705	1.1228	.05153	.009	51,586
Japanese (90)	Japan	1583	1.0790	.02995	.004	168,790
Shanghai (23-35)	China	730	1.0317	.03914	.209	unknown

Note for Table 2:

(i) * This amount is the average total pool size including all pools.

(ii) p-value is associated with the following pair of hypotheses : H_0 : $\beta=1$ and H_1 : $\beta>1$ based on the test statistic : $(\hat{\beta}-1)/\hat{se}(\hat{\beta})$.

VI. Comparison with other models

1. Ali's implicit estimates of winning probabilities :

Ali (1977) assumed expected utility maximizing bettors whose betting opportunity is limited to one race.

As before, let π_i= winning probability of horse i and O_i= odds on horse i. Ali computed the weighted averages of all the realized odds $\overline{O}_i$ and approximated the winning probabilities of horses ($\hat{\pi}_i$) in different favourite positions by the relative winning frequencies. He estimated the following utility of wealth function by least squares (for details of the derivation, see Ali (1977)) :

$$\log_{10}U(x) = -1.7202 + 1.1784\log_{10}(x) \quad (5)$$

$$R^2 = 0.9981$$

(The estimated intercept term wasnot reportedcorrectlyinhispaper.)

or,

$$U(x) = 0.01905\ x^{1.1784} \quad (6)$$

where x = wealth = $1+O_i$ for any horse i.

By evaluating an absolute risk-aversion measure, Ali concluded that the representative bettor takes more risk as his capital declines. Using the utility and wealth relationship in (6), we can estimate the winning probabilities for each race as :

$$\hat{\pi}_i = \frac{1/U(1+O_i)}{\sum_j 1/U(1+O_j)} \quad (7)$$

where $U(1+O_i)$ are computed using (6) and the observed odds. To compare these estimates with those of our simple constant-β model, we compute the log likelihood and the results are shown in Table 3.

Table 3

Comparison among different methods of estimation

Method	log lik	#parameters
Constant-β model	-26834.42	1
Ali's estimates	-26840.42	2
Win bet fractions	-26895.51	0

Because the models are not nested, likelihood ratio tests are not possible, but by directly comparing the log likelihood values, our constant-β model is slightly better than the model implied by Ali's utility function.

The following formal test, based on log likelihoods, should be more accurate. Cox's test (see Cox (1961,62)) is carried out to test H_f : constant-β model vs. H_g : Ali's estimates.

The test involves two stages. First, evaluate :

$$T_f = \frac{(loglik_f - loglik_g) - E(loglik_f - loglik_g \mid H_f)}{SD(loglik_f - loglik_g \mid H_f)}$$

then, evaluate :

$$T_g = \frac{(loglik_g - loglik_f) - E(loglik_g - loglik_f \mid H_g)}{SD(loglik_g - loglik_g \mid H_g)}$$

where $loglik_i$ = the log lik under H_i with the estimated parameter (i=f,g),

$E(.\mid H_i)$ and $SD(.\mid H_i)$ are expectation and standard deviation when H_i is true.

Under H_f, $T_f \sim N(0,1)$. Under H_g, $T_g \sim N(0,1)$.
If H_g is better than H_f, T_f will be largely negative.
If H_f is better than H_g, T_g will be largely negative.

It can be difficult to evaluate the expectations and the standard deviations. Some authors (e.g. Li(1989)) propose using a bootstrap method (Efron(1982)).

In our case, 100 bootstrapped samples are used and the following test statistics are obtained :

$T_f = 3.0521$,
$T_g = -5.0218$.
Therefore, the constant-β model appears preferable at any reasonable significance level.

2. AMQ's logit model :

Asch, Malkiel and Quandt (1984) fitted a logit model (hereafter called AMQ's logit model) for the relationship between the true winning probability and the bet fraction. AMQ's logit model is as follows :

$$\pi_i = \frac{\exp(qP_i)}{\sum_j \exp(qP_j)} \quad \text{where } q \text{ is a parameter} \qquad (10)$$

(Note that in their original paper, the parameter is β, we change this to q in order to avoid confusion.)

Using their Atlantic City data (712 races), we compare the constant-β model with the AMQ logit model. The results are shown in Table 4.

By comparison of log likelihoods, AMQ's logit model ranks below the constant-β model and even the simplest model $\pi_i = P_i$. Note that in the constant-β model, when β=1, it reduces to the simplest model $\pi_i = P_i$ but the AMQ logit model cannot be reduced to this by restricting the parameter q.

Table 4

Comparisons among different models

Method	loglik	est parameter
Constant-β model	-1233.99	$\hat{\beta}$=1.1023
Win bet fractions	-1235.53	———
AMQ's logit model	-1264.56	$\hat{q}$=6.6588

(Note that our result for AMQ's logit model is very close to but not exactly the same as the result reported in their original paper (log likelihood = -1265). They used :

$P_i \simeq (1-t)/(1+O_i)$ so $\sum P_j \neq 1$.

Results reported here use :

$P_i = \frac{1}{1+O_i} / \sum_j \frac{1}{1+O_j}$ so that $\sum P_j = 1$.)

Cox's test is again carried out to test H_f : $\pi_i = P_i$ vs. H_g : AMQ's logit model. The number of bootstrapped samples = 100. The following values of test statistics are obtained :

$T_f = 0.0034$,

$T_g = -8.2158$.

Therefore, we can reject (by Cox's test) AMQ's logit model in favour of the model $\pi_i = P_i$ at any reasonable level of significance. Using a likelihood ratio test, the $\pi_i = P_i$ model can only be rejected in favour of the constant-β model at 10 % level of significance.

VII. Conclusions

Our task has been to provide an improved statistical analysis of betting data. After going through a relatively refined modelling process, the constant-β model seems to have appealing characteristics. It provides slightly better results than Ali's implicit estimates of winning probabilities and is much better than AMQ's logit model.

Under this model, β is found to be significantly greater than one at some racetracks, which provides support for previous findings of a favourite-longshot bias. In other racetracks, however, this bias cannot be found. We find no reason to support any strong conclusion of a generalized risk preference among gamblers.

References

Aitchison,J. (1986) The statistical analysis of compositional data. Chapman and Hall.

Ali, M.M. (1977) "Probability and utility estimates for racetrack bettors." *Journal of Political Economy* 84 (August) :803-15.

Asch,Peter;Malkiel,B. and Quandt,R. (1982) "Racetrack betting and informed behavior." *Journal of Financial Economics* 10 (June) : 187-94.

Asch,Peter;Malkiel,B. and Quandt,R. (1984) "Market efficiency in racetrack betting." *Journal of Business* 57 (April) : 165-74.

Busche,K. and Hall,C.D. (1988) "An exception to the risk preference anomaly." *Journal of Business* 61 : 337-46.

Cox, D.R. (1961) "Tests of separate families of hypotheses." *Proc. 4th Berkeley Symp.*, 1 : 105-123.

Cox, D.R. (1962) "Further results on tests of separate families of hypotheses." *Journal of Royal Statistical Society B* 24, No.2 : 406-424.

Efron, B. (1982) The Jackknife, the Bootstrap and other Resampling Plans. Philadelphia : Society for Industrial and Applied Mathematics.

Hausch,D.B.;Ziemba,W.T. and Rubinstein,M. (1981) "Efficiency of the market for racetrack betting." *Management Science* 27 (December) : 1435-52.

Hosmer, D.W. and Lemeshow S. (1989) Applied logistic regression. John Wiley & Sons.

Judge, G.G.; Griffiths, W.E.; Hill, R.C.; Lutkepohl, H. and Lee, T.C. (1985) The theory and practice of Econometrics. 2nd edition. John Wiley & Sons.

Li, C.K.T. (1986) Statistical problems involving compositions in a covariate role. M.Phil. thesis, Department of Statistics, University of Hong Kong.

Li, W.K. (1989) "Discriminating nonlinear time series models using a bootstrapped Cox test." technical report, Department of Statistics, University of Hong Kong.

Sakamoto,Y.; Ishiguro, M. and Kitagawa, G. (1986) Akaike information criterion statistics. D.Reidel publishing co..

Snyder, Wayne W. (1978) "Horse racing : Testing the efficient markets model." *Journal of Finance* 33 (September) : 1109-1118.

Thaler, R.H. and Ziemba, W.T. (1988) "Anomalies parimutuel betting markets : racetracks and lotteries." *Journal of Economic Perspectives* 2, No.2 (Spring) : 161-174.

Fig. 1

A-class models for 6-horse-races
(no. of races = 807)

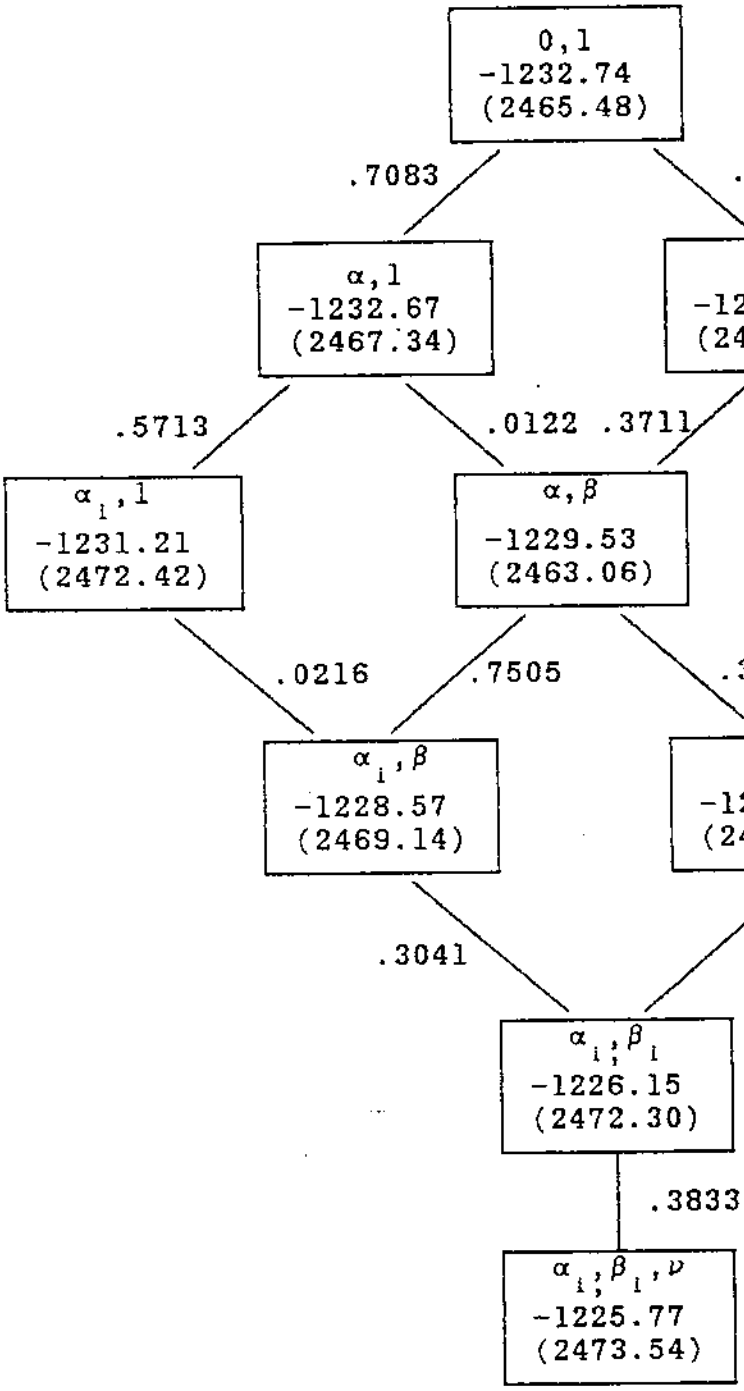

Fig. 2

A-class models for 7-horse-races
(no. of races = 3532)

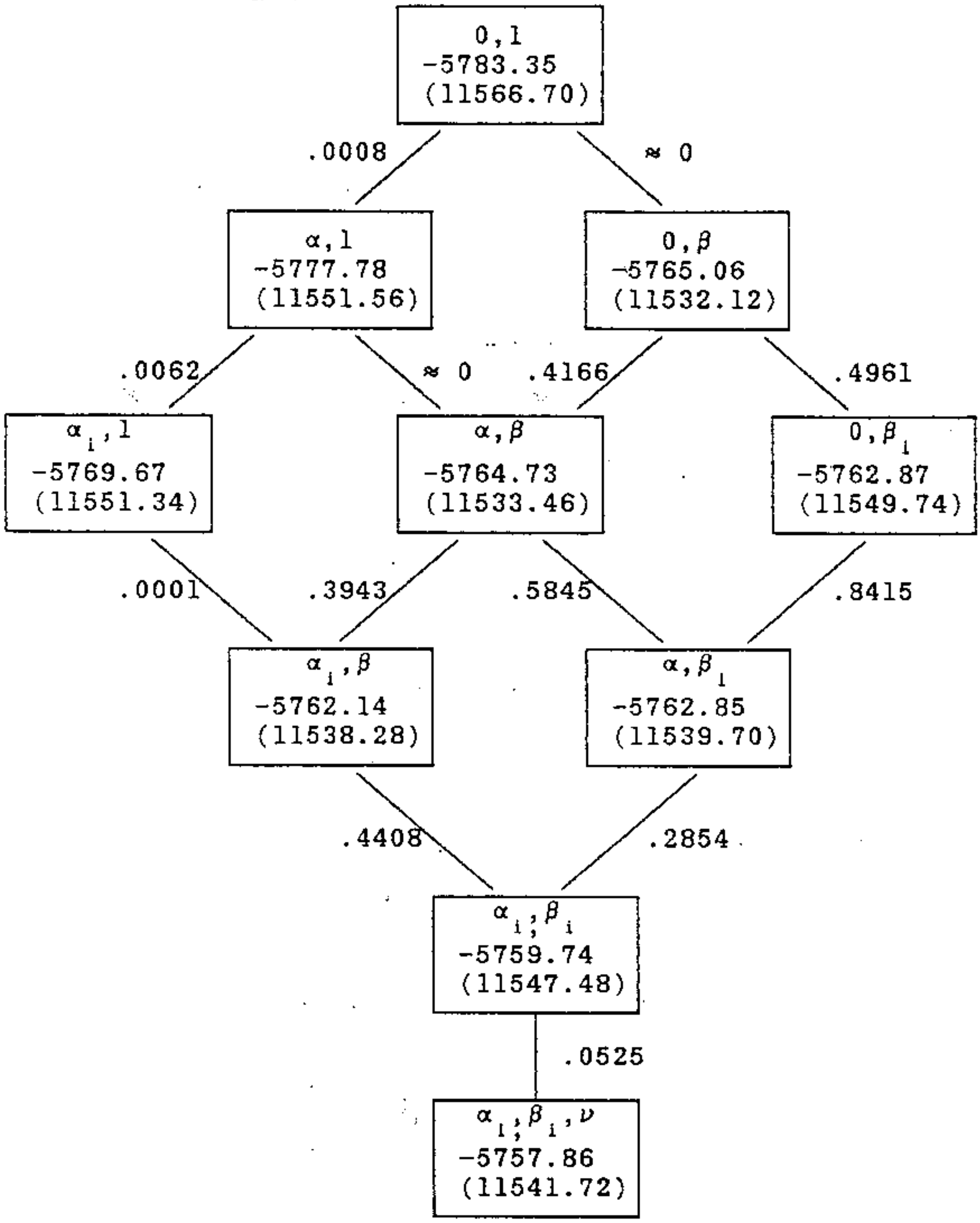

Fig. 3

A-class models for 8-horse-races
(no. of races = 15450)

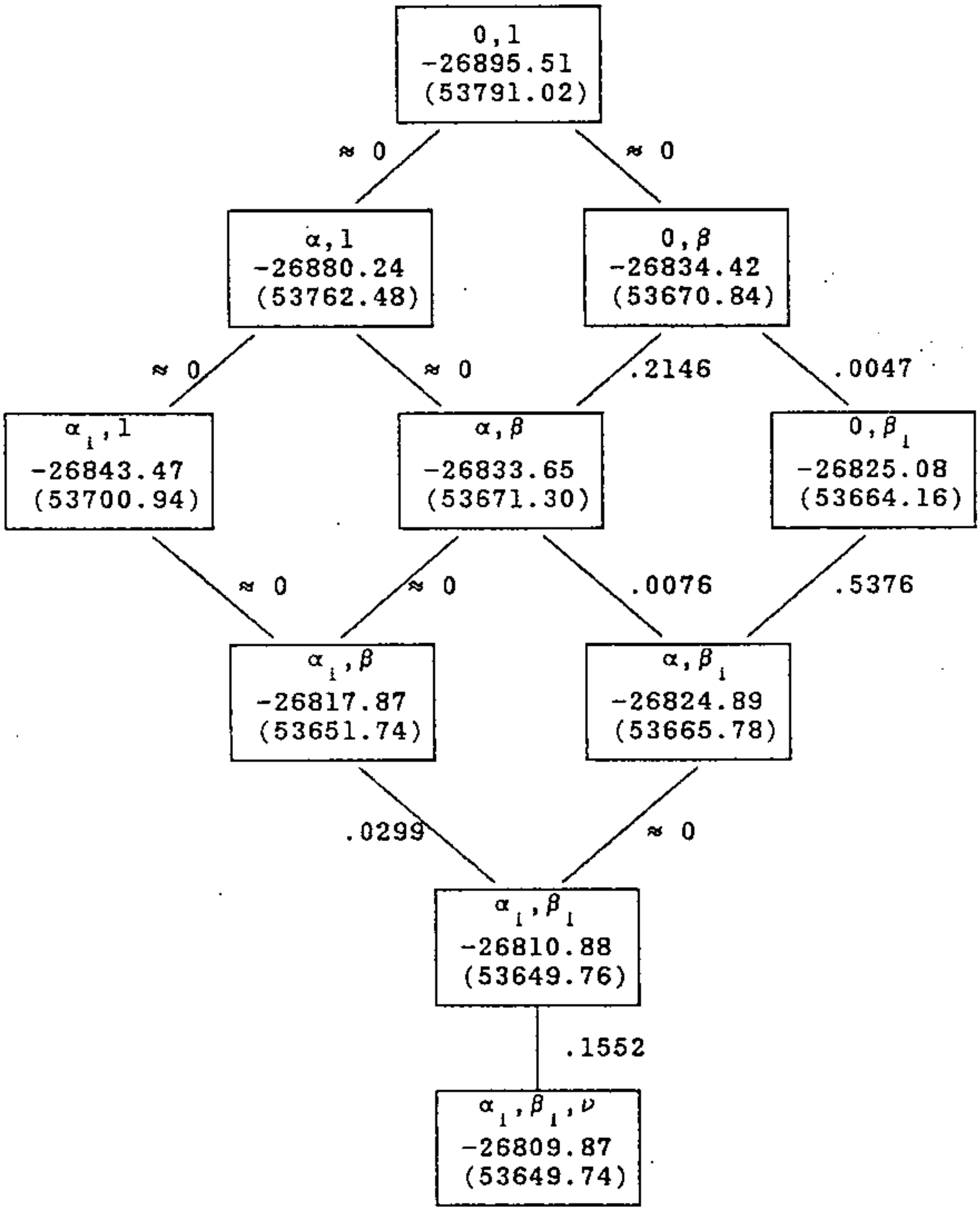

Fig. 4

L-class models for 6-horse-races
(no. of races = 807)

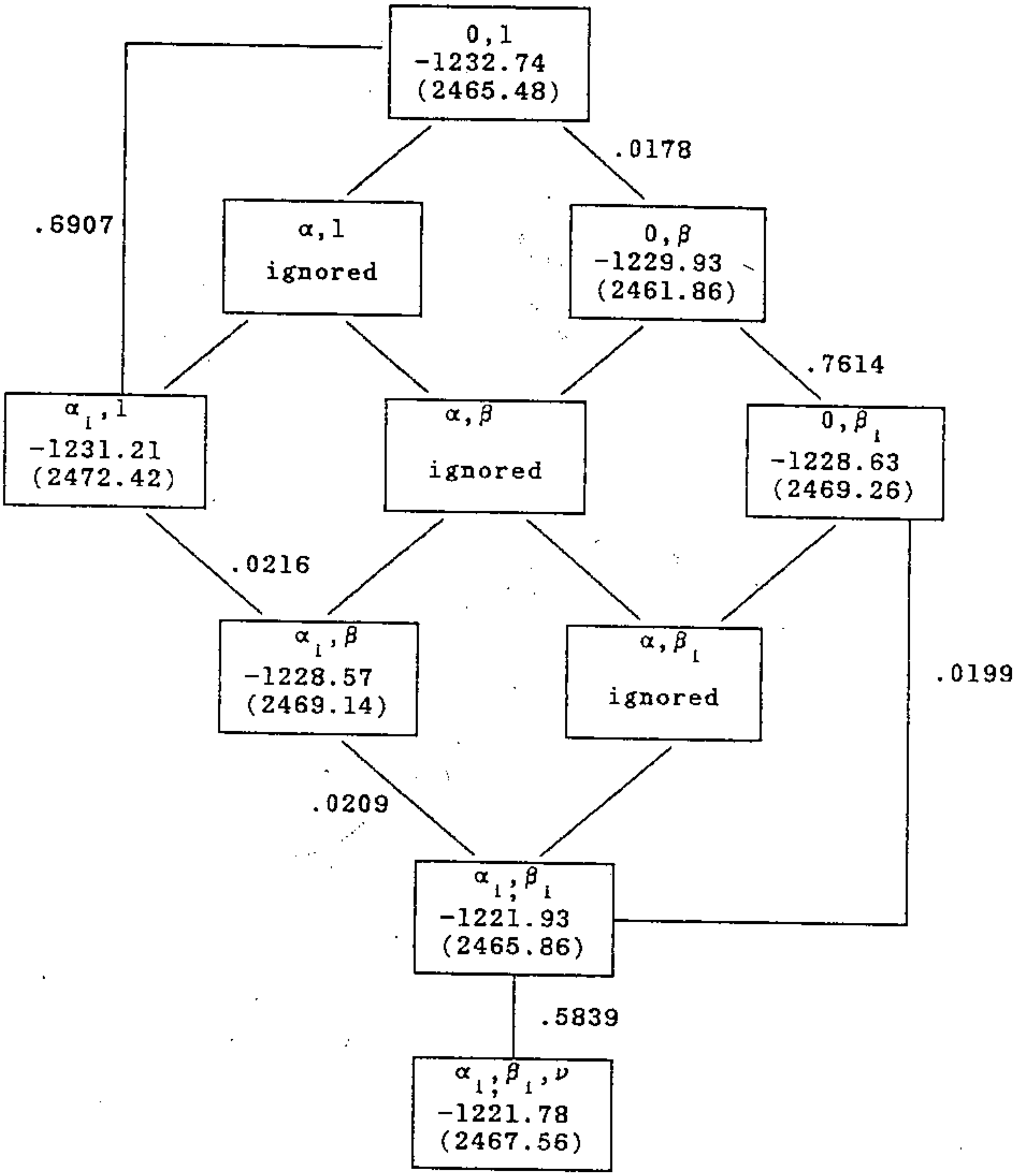

Fig. 5

L-class models for 7-horse-races
(no. of races = 3532)

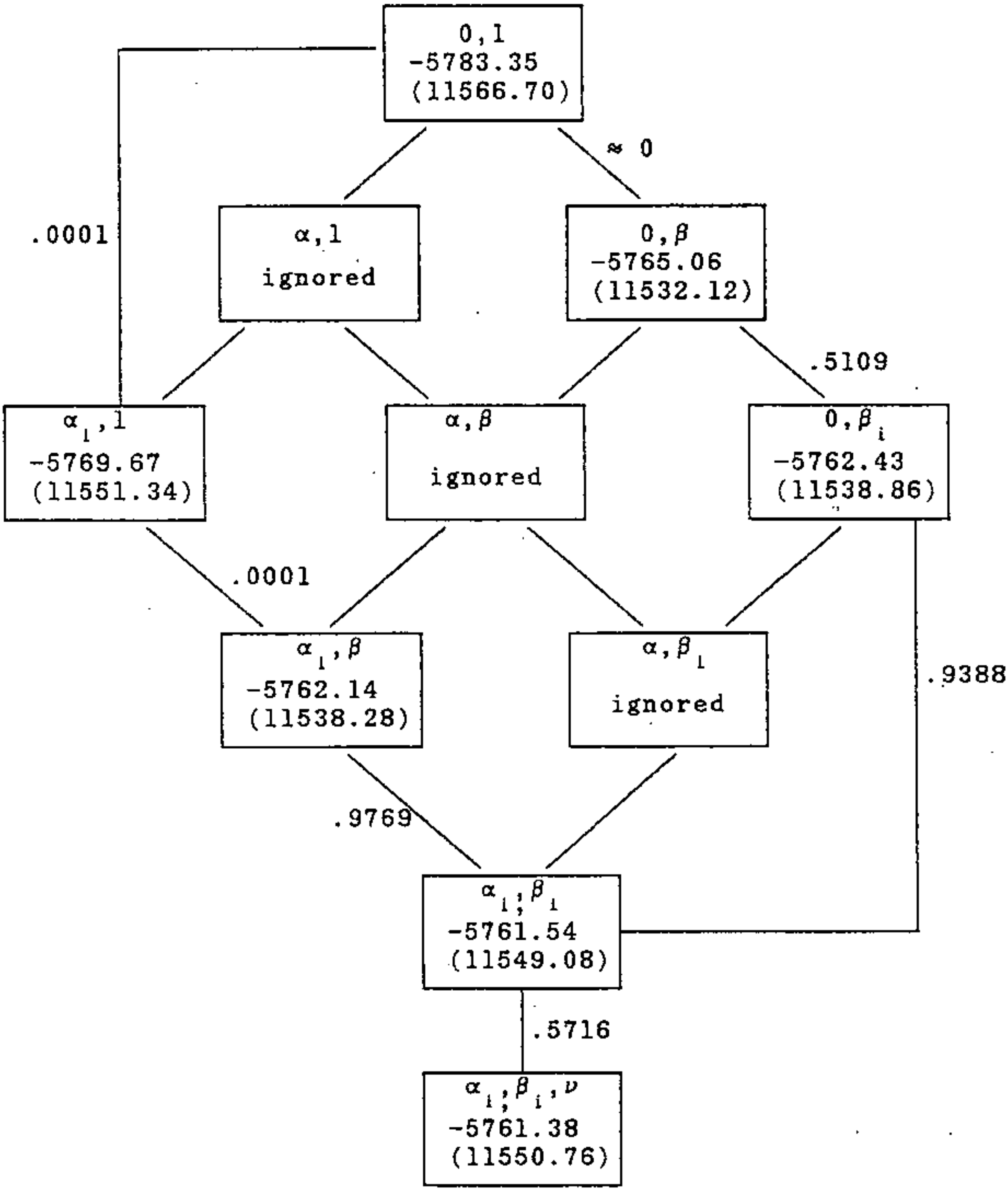

Fig. 6

L-class models for 8-horse-races
(no. of races = 15450)