

더러브렛 경주마의 주파기록에 대한 유전모수 추정과 연도별 개량량 분석

오승윤* · 박종은** · 이정란*** · 이진우**** · 오희석*** · 김희발**

서울대학교 협동과정 생물정보학전공*, 서울대학교 농생명공학부**, 서울대학교 통계학과***, 한국마사회****

Estimation of Genetic Parameters and Annual Trends for Racing Times of Thoroughbred Racehorses

Seungyoon Oh*, Jong Eun Park**, Jeong Ran Lee***, Jin Woo Lee****, Hee Seok Oh*** and Hee-bal Kim**

Interdisciplinary Program in Bioinformatics, Seoul National University*, Department of Agricultural Biotechnology, Seoul National University**, Department of Statistics, Seoul National University***, Korea Racing Authority****

ABSTRACT

The aim of this study was to estimate genetic parameters and annual trends on the racing performance of Thoroughbred horses by a statistical analysis of the resulting records. We used the racing results of 245,979 observations for 13,458 horses recorded in 19 years of race held at Seoul and Busan racing tracks, provided by Korea Racing Authority. After a careful adjustment of some variables such as racing times, jockey and trainer numbers and the average prize a horse won, we selected significant factors that explain the result of racing records of a horse by stepwise AIC and BIC methods. The estimated heritability and repeatability were 0.322 and 0.332, respectively. The average of annual phenotypic and genetic improvement was -0.166 seconds and -0.161 seconds, respectively. Based on the statistical approach, we established reasonable animal model of well-set variables, which is important in the study on estimating performance of racing horses.

(Key words : Thoroughbred, Heritability, Repeatability, Genetic trend)

I. 서 론

유럽에서 시작된 경마는 전세계적으로 확산되어 오늘날 세계 150개 이상의 많은 국가에서 경기가 개최되고 있으며 경제적인 측면에서도 대규모의 이윤을 창출하는 하나의 산업으로 발전하고 있다. 국내의 경마산업은 한국마사회 중심의 선도 하에 점차 대중화되어 2007년 기준 약 6조 5천억원의 매출규모를 가진 큰 산업으로 성장하였고, 일반시민의 건전한 여가생활이자 행정자치단체의 중요한 재원으로 인식되고 있다.

이처럼 대규모화되며 높은 부가가치를 창출하고 있는 국내의 경마산업은 이제 양적, 질적 측면에서 성공적인 국제화를 목표로 하여 이를 위한 다양한 시도가 이루어지고 있다. 이 과정에서 국내 경마산업이 세계 수준의 강한 경쟁력을 갖추기 위한 요건들 중 마필산업 육성에 대한 중요성이 대두되고 있으며, 실제로 우수한 경주마의 생산과 육성 부문에 대해 국가차원의 중장기적 계획들이 수립되어 있는 상황이다.

따라서 본 연구는 한국 경마산업의 발전에 있어서 중요시되는 수준 높은 마필산업의 구축에 있어 기반이 되는

국내 경주마들의 유전능력과 연도별 개량량에 대한 분석을 실시하였다. 국내에서 이루어진 19년간의 경기 결과를 바탕으로 한 유전능력의 평가와 개량 추세의 분석은 현재 국내 경마산업의 위치를 확인하고 더 높은 세계적 수준으로의 도약에 있어 나아가야 할 방향을 제시하는 지표로 활용될 수 있다.

II. 재료 및 방법

1. 공시재료

본 연구의 원자료는 한국마사회로부터 제공받았으며 서울과 부산경마공원에서 1990년부터 2008년까지 19년 동안 수집되어온 더러브렛 품종의 경주마 13,458두의 245,979개의 관측치가 분석에 사용되었다. 관측치의 분포는 Table 1과 같으며, 총 9대에 걸친 66,644개의 종모마, 종빈마와 자손간 혈연 관계를 나타내는 가계도 정보가 분석에 이용되었다. 분석에 사용된 총 13,458두의 경주마는 원산지에 따라 한국산 2,766두, 외국산 10,692두로 분류할 수 있으며, 성별은 암말 6,944두, 수말 3,510두, 거세마 3,004두로

Corresponding author : Hee-bal Kim, Department of Agricultural Biotechnology Seoul National University San 56-1, Sillim-Dong, Gwanak-Gu Seoul 151-742, KOREA
Tel: 82-2-880-4803, Fax: 82-2-883-8812, E-mail: hee-bal@snu.ac.kr

Table 1. Number of records and horses by race year

Year	No. of Records			No. of Horses ¹⁾		
	Seoul	Busan	Total	Seoul	Busan	Total
1990	12,225	—	12,225	1,543	—	1,543
1991	11,209	—	11,209	1,538	—	1,538
1992	8,968	—	8,968	1,594	—	1,594
1993	9,782	—	9,782	1,584	—	1,584
1994	10,572	—	10,572	1,437	—	1,437
1995	11,288	—	11,288	1,561	—	1,561
1996	11,589	—	11,589	1,573	—	1,573
1997	11,999	—	11,999	1,628	—	1,628
1998	12,441	—	12,441	1,536	—	1,536
1999	12,261	—	12,261	1,640	—	1,640
2000	12,270	—	12,270	1,589	—	1,589
2001	12,079	—	12,079	1,686	—	1,686
2002	11,870	—	11,870	1,708	—	1,708
2003	11,855	—	11,855	1,722	—	1,722
2004	11,539	297	11,836	1,661	252	1,913
2005	11,514	4,517	16,031	1,734	735	2,469
2006	11,875	6,403	18,278	1,890	954	2,844
2007	12,433	6,683	19,116	1,963	1,011	2,974
2008	12,508	7,802	20,310	1,907	1,133	3,040
Total	220,277	25,702	245,979	11,477	1,981	13,458

¹⁾ Unique number of horses.

분류할 수 있다. 245,979개 관측치의 원산지, 성별, 연령에 대한 정보는 Table 2에 제시하였다.

2. 통계적 방법

(1) 자료의 전처리와 변수값들의 보정

각 변수들이 실제 경주에서 가지는 의미에 비추어, 비합리적인 관측값을 제외 또는 보정하였다. 그리고 반응변수인 모든 관측치의 주파기록을 동일한 경주거리에 대한 주파기록으로 변환하였다. 설명변수 중 기수 번호, 조교사 번호는 수득평균상금에 따라 순위화한 후, 이 순위화된 값을 변수값으로 지정하였다.

(2) 설명변수의 선택과 개체모형의 수립

주파기록에 대한 후보 설명변수들을 대상으로 AIC (Akaike, 1973)와 BIC (Schwarz, 1978)를 이용한 단계적 변수 선별법을 사용하였다. 후보 설명변수가 범주 개수가 많은 이산형 변수인 경우에는 일원분산분석과 Tukey (1953)와 Montgomery (2005)에 제안된 Tukey's HSD (honestly significant difference) 검정을 사용하여 범주의 수준간에 차이가 있는 변수에 대해 어느 부분에서 차이가 났는지 확인하였다. 이 과정들은 R 프로그램을 이용하였

다. 개체모형은 다음과 같으며, 개체모형 방정식의 해와 개별 경주마들의 육종가는 Wombat (Meyer, 2007) 프로그램의 DF-REML을 이용하여 추정하였다. 이 과정에서 경주마들의 가계도 정보를 이용하여 육종가에 종모마와 종빈마의 효과가 적용되도록 하였으며, 분산성분 값은 SAS 프로그램의 mixed linear model을 이용하여 산출하였다.

$$y = Xb + Za + R_1yr + R_2m + R_3j + R_4mg + e$$

위에서,

y = 주파기록에 대한 관측치 벡터

b = 출주번호, 생산국, 주로상태, 날씨, 별정경주의 구분, 별정경주의 종류, 성별, 경주의 구분, 부담의 구분, 경마장의 고정효과에 대한 벡터 (n=14, 14, 8, 5, 7, 4, 3, 3, 3, 2)

a = 유전적 상가효과에 대한 벡터 (n=67,841)

yr = 마령의 임의효과에 대한 벡터 (n=13)

m = 주로 함수율의 임의효과에 대한 벡터 (n=19)

j = 기수의 임의효과에 대한 벡터 (n=273)

mg= 조교사의 임의효과에 대한 벡터 (n=123)

X, Z, R₁, R₂, R₃, R₄ = 각 효과의 계수행렬

e = 임의 오차효과에 대한 벡터

Table 2. Characteristics of records by race year

Year	Country of birth		Sex			Average age(yr)
	Domestic	Foreign	Male	Female	Gelding	
1990	3,844	8,381	2,216	6,059	3,950	5.4
1991	5,443	5,766	1,888	5,450	3,871	5.5
1992	4,920	4,048	1,513	4,362	3,093	5.4
1993	5,223	4,559	1,610	4,829	3,343	5.1
1994	5,184	5,388	1,543	5,415	3,614	5.0
1995	5,521	5,767	1,207	6,079	4,002	4.9
1996	5,697	5,892	1,093	6,315	4,181	4.9
1997	4,958	7,041	1,177	6,420	4,402	4.8
1998	3,979	8,462	1,269	6,633	4,539	4.9
1999	4,225	8,036	1,649	6,656	3,956	4.7
2000	3,809	8,461	2,378	6,557	3,335	4.5
2001	2,788	9,291	3,054	6,460	2,565	4.2
2002	1,184	10,686	3,423	6,132	2,315	4.0
2003	543	11,312	3,725	5,916	2,214	3.9
2004	459	11,377	3,871	5,685	2,280	3.8
2005	706	15,325	5,641	7,926	2,464	3.5
2006	734	17,544	6,742	8,909	2,627	3.6
2007	576	18,540	7,260	8,470	3,386	3.6
2008	509	19,801	7,486	8,845	3,979	3.6

III. 결과 및 고찰

주파기록이 전체 범위의 99.9% 이상에 해당하는 관측치들은 정상적인 경주가 이루어진 것으로 인정할 수 없어 이상치로 간주하고 제외하였다. 그리고 주로 함수율이 20%가 넘는 관측치들의 주로상태 변수 값을 불량으로 지정하였다. 또한 모래 두께가 0인 관측치들은 정상적인 기록으로 인정할 수 없어 제외하였다. 수득상금은 그 해의 연평균 경주상금으로 나누어 주어 과거와 현재의 상금규모 차이에서 발생하는 오차를 제거하였다. 이러한 전처리 과정들을 거친, 실제 이용된 최종 관측치의 개수는 총 245,979개였다.

주파기록은 경주거리가 늘어남에 따라 증가하므로 수립된 개체모형에 모든 관측치의 정보를 이용함과 동시에 모든 주파기록 사이의 절대적인 비교가 가능하도록 보정하기 위하여 기준이 되는 하나의 경주거리 설정이 요구되었으며, 이에 대해 관측치의 개수가 많고 실제 경마에서 가장 관심이 집중되는 종목인 1,400 m를 기준 경주거리로 정하였다. 기준 경주거리와 각 경주거리의 평균주파기록 차이만큼을 해당 경주거리의 개별 관측치들의 주파기록에 가감을 통해 보정하였고 보정 전후 전체 주파기록 분포의 변화는 Fig. 1과 같으며 보정된 주파기록의 연도별 요약 통계치는 Table 3와 같다.

이산형 변수의 형태인 기수 번호, 조교사 번호는 그들의 평균수득상금에 따라 순위를 매기고 새로 매겨진 순위 값을 변수값으로 지정하였다. 이 과정을 통해 변수의 성격을 주파기록과 연관성을 가지는 연속형 변수로 변환시키면서, 향후 매년 새로운 유입이 발생할 기수와 조교사 정보에 대해 논리적으로 유연한 대처가 가능하도록 하였다. 보정 후의 기수 번호, 조교사 번호에 따른 평균수득상금의 분포는 Fig. 2와 같다.

주파기록에 대한 후보 설명변수의 특성에 따라 범주 개수가 많은 이산형 변수, 범주 개수가 적은 이산형 변수 그리고 연속형 변수라는 세 가지 그룹으로 구분하였다. 범주 개수가 많은 이산형 변수에는 출주번호, 생산국, 주로상태, 날씨, 별정경주의 구분, 별정경주의 종류, 경주번호, 군, 출주두수가 해당하였고, 범주 개수가 적은 이산형 변수에는 성별, 경주의 구분, 부담의 구분, 경마장이 해당하였으며, 마령, 주로 함수율, 기수, 조교사, 부담중량, 모래 두께는 연속형 변수에 속하였다.

범주 개수가 많은 이산형 변수에 대해서는 일원분산분석과 Tukey's HSD 검정을 사용하여 선형 모형의 예측력을 높이고자 가변수의 개수를 최대한 줄이려 하였지만 유의수준 0.05에서 뚜렷하게 재그룹 되는 변수는 없었다. AIC, BIC를 이용한 단계적 변수 선발을 통해 출주번호, 생산국, 주로상태, 날씨, 별정경주의 구분, 별정경주의 중

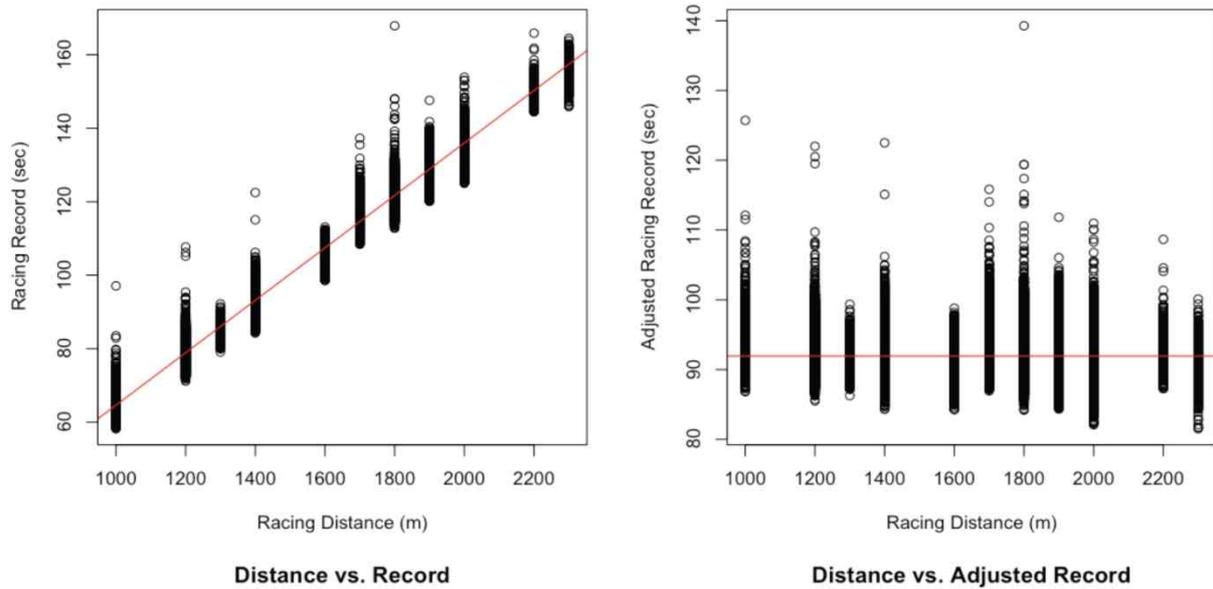


Fig. 1. Distribution of racing time and new distribution of adjusted racing time.

Table 3. Minimum, 1st quartile, median, mean, 3rd quartile and maximum of adjusted racing time by race year

Year	Summary statistics (sec)					
	Min	1st Qu	Median	Mean	3rd Qu	Max
1990	88.43	93.21	94.24	94.34	95.38	105.04
1991	89.40	93.71	94.72	94.83	95.82	119.38
1992	87.26	93.11	94.27	94.36	95.51	113.78
1993	88.06	93.24	94.23	94.38	95.34	111.83
1994	87.46	92.91	94.01	94.08	95.18	125.72
1995	84.71	91.88	93.30	93.23	94.60	108.31
1996	86.41	92.71	93.92	94.04	95.22	119.38
1997	87.21	92.71	94.00	94.09	95.34	139.28
1998	86.50	92.22	93.54	93.60	94.91	109.71
1999	86.50	92.61	93.82	93.93	95.12	115.84
2000	87.40	92.67	93.98	94.11	95.37	115.08
2001	86.13	92.30	93.41	93.50	94.61	110.07
2002	86.90	91.80	93.02	93.12	94.32	122.01
2003	84.77	90.92	92.30	92.29	93.61	120.51
2004	83.37	90.91	92.32	92.33	93.71	107.68
2005	81.51	90.92	92.23	92.23	93.52	100.93
2006	83.57	90.50	91.90	91.88	93.24	102.78
2007	83.77	90.21	91.64	91.70	93.10	103.34
2008	82.07	90.38	91.71	91.72	93.10	102.48

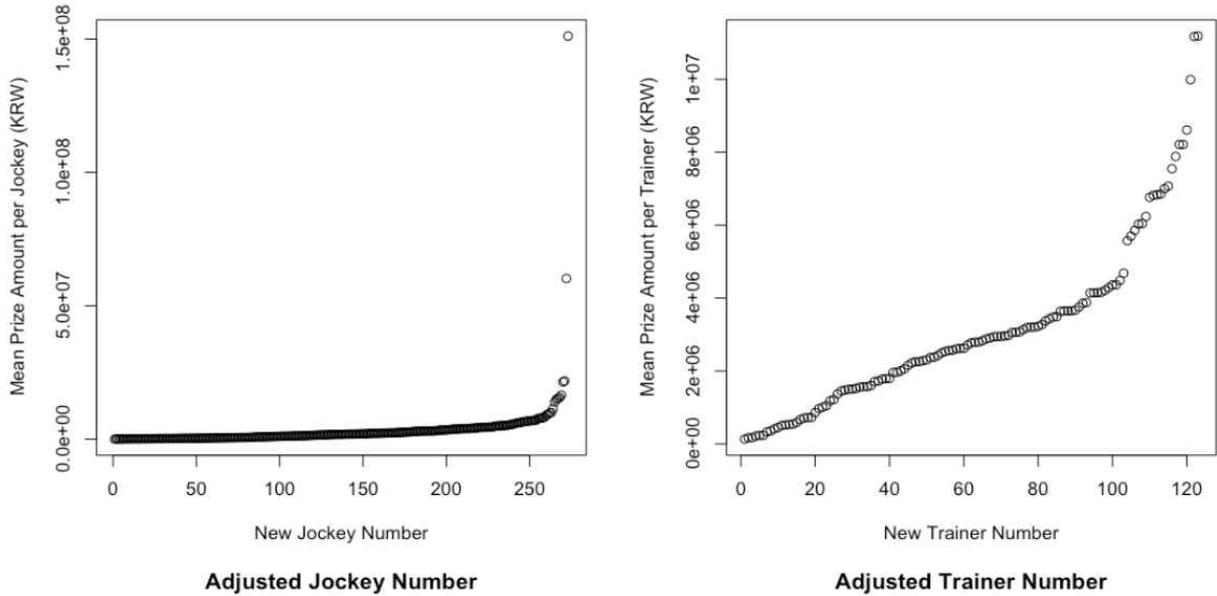


Fig. 2. Distribution of average prize income by new jockey number and new trainer.

류, 성별, 경주의 구분, 부담의 구분, 경마장, 마령, 주요 함수율, 기수, 조교사가 유의미한 설명변수로 나타났으며, 이 결과는 AIC와 BIC에서 모두 동일하였다.

국내에서 활동한 경주마들의 주과기록에 대한 유전력과 반복력은 각각 0.322와 0.332로 추정되었으며, 이는 선발에 의한 개량이 가능한 수준인 중도의 유전력을 의미한다. 개체모형의 분산성분과 유전모수는 Table 4에 제시하였으며 협의의 유전력 개념을 사용하였다. 영구환경효과는 동일 개체의 반복된 기록에 꾸준한 영향력을 미칠 수 있는 항목이어야 하므로 한 경주마의 성장에 있어서의 영양관리, 훈련방식 등은 고정된 담당 조교사의 관리하에 이루어진다는 점을 감안한다면 담당 조교사의 효과가 해당 경주마의 경주능력에 있어 근본적이고 영구적인 영향을 미치는 중요 요소라고 판단할수 있어 조교사 효과를 영구환경효과로 사용하였다.

Bakhtiari와 Kashan (2009)의 보고에 따르면 이란의 경주마들의 유전력과 반복력은 0.11, 0.19로 추정되었으며, Mota 등 (2005)의 연구에서는 브라질의 경주마들은 0.06, 0.20의 유전력과 반복력을 나타내었다. 또한 Chico (1994)에 따르면 스페인의 경주마들은 0.01의 매우 낮은 유전력과 반복력을 나타내어, 이들과 비교하였을 때 국내의 추정 값이 매우 높았다. 반면, Ekiz와 Kocak (2007)의 연구에서는 터키경마장에서의 기록을 이용하여 0.265, 0.363의

유전력과 반복력을 추정하였으며, Moritsu 등 (1998)에 의하면 일본에서 활동하는 경주마들을 연구한 결과 잔디와 진흙 주로에서 각각 0.29, 0.18의 유전력이 추정되어 국내의 결과와 비슷하거나 국내가 다소 높았다.

주과기록에 대한 연평균 표현형 개량량과 유전적 개량량은 -0.166 초, -0.161 초로 산출되었으며, 결정계수는 각각 0.8371, 0.9558로 높은 집합도를 보였다. 19년간의 개량 추세는 Fig. 3와 같다. 연평균 표현형 개량량과 유전적 개량량 모두 90년대에는 미세하게 기록이 단축되다가 2000년대 들어서 급격하게 단축되는 추세를 나타냈다. 연평균 유전적 개량량은 Wombat 프로그램에서 산출된 각 경주마들의 육종가를 이용하여 도출한 연도별 평균 육종가의 변화 추세이며, 이는 Park 등 (2008)의 보고한 -0.027 초와 매우 큰 차이를 보였다. 2000년대 들어서 급격하게 향상된 주과기록의 상당 부분은 유전적 개량의 영향을 받은 것으로 생각된다.

이러한 경주마의 관측기록으로부터 유전모수와 개량량을 추정하는 연구에 있어서, 수립한 개체모형의 설명력이 추정의 정확도를 결정짓는 중요 요인이라고 생각된다. 이를 위하여 충분한 관측치 정보를 바탕으로 한, 경주마의 능력을 효과적으로 대표할 수 있는 반응변수의 설정과 반응변수를 유의미하게 설명하는 변수들의 선발이 이루어져야 하며, 설명변수의 선발은 통계학적 기법에 기반한 체

Table 4. Variance component and genetic parameter estimates of racing times

σ_g^2	σ_{ye}^2	σ_m^2	σ_j^2	$\sigma_{pe(mg)}^2$	σ_e^2	σ_p^2	h^2	r
1.522	0.201	0.022	0.077	0.051	2.861	4.734	0.322	0.332

σ_g^2 : additive genetic variance, σ_{ye}^2 : age (year) variance, σ_m^2 : ground moisture variance, σ_j^2 : jockey variance, $\sigma_{pe(mg)}^2$: permanent environmental variance (trainer variance), σ_e^2 : residual variance, σ_p^2 : phenotypic variance, h^2 : heritability, r: repeatability

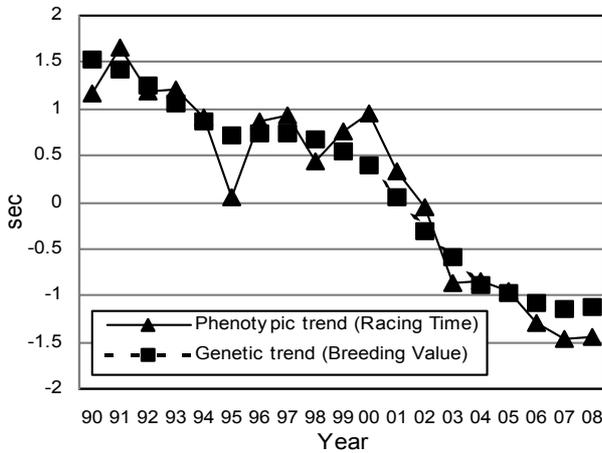


Fig. 3. Phenotypic and genetic trends over the year.

계적인 과정이 되어야 한다. 또한 각 변수들은 실제 경마의 특성을 잘 반영함에 동시에 선형 개체모형의 예측력 저하 방지를 위해 최소한의 범주만을 포함하도록 하는 보정이 필요하다.

이번 연구에 사용된 경주마의 주과기록을 기준으로 한 유전능력 평가 방법은 모든 관측치 정보를 이용하기 위해 선 하나의 기준거리에 대한 표준화된 기록으로 보정하는 과정을 거쳐야 하며, 경기내의 상황과 전략 등의 유동적인 환경 조건들을 정밀하게 반영하기 어려운 단점도 존재한다. 경주거리가 증가함에 따라 주과시간이 증가하므로 모든 관측치의 경주기록간의 절대적인 비교가 불가능하여 각 경주거리를 명확하게 구분하여 분석하는 방법과 모든 관측치를 한번에 분석하는 방법 사이의 딜레마가 존재하며, 실제 경주의 상황과 기수의 전략에 따라 출발선부터 도착점까지 구간별로 속도가 변화할 수 있는 점을 효과적으로 반영하기 어렵기 때문이다.

따라서, 주과기록 이외의 경주마의 경주능력을 설명해 줄 수 있는 착순, 수득상금과 같은 변수들의 반응변수로의 도입을 통해 주과기록만으로 완벽하게 설명할 수 없는 부분을 보완해야 할 것이다. 특히, 경주능력을 설명할 수 있는 여러 변수들에 대한 주성분분석을 통해 각 변수들의 설명력이 조합된 새로운 변수를 생성하여 활용하는 방법을 고려해볼 수 있다. 설명변수에 있어서는 실제 경주상황과 환경을 더욱 정확하게 반영하도록 하는 새로운 변수의 도입을 통해 모형의 설명력을 높일 수 있을 것이다.

IV. 요약

본 연구는 1990년부터 2008년까지 서울경마공원과 부산경마공원에서 활동한 경주마 13,458두의 총 245,979개의 경주관측기록을 이용하여 국내 더러브렛 경주마의 유전능력 및 연도별 개량량을 분석하였다. 모든 관측치들의 주과기록을 1400m 경주거리에 맞추어 보정하고, 기수와 조교사에 대해 평균수득상금의 순위에 따라 번호를 부여하였

으며, 후보 설명변수들에 대해 AIC, BIC를 사용한 단계적 변수 선별법을 적용하였다. 19년간 국내에서 활동한 경주마들의 주과기록에 대한 유전력과 반복력은 각각 0.322, 0.332로 추정되었고, 연평균 표현형 개량량과 유전적 개량량은 -0.166 초, -0.161 초로 산출되었다. 경주마의 유전모수와 개량량의 추정 연구에 있어 무엇보다 설명력 높은 개체모형의 수립이 중요하다고 생각되며, 이는 실제 경주마의 능력을 효과적으로 반영하는 반응변수의 설정과 통계학적 기법을 활용한 체계적인 설명변수의 선별을 통해 가능하다.

V. 사 사

본 연구는 한국마사회의 지원 하에 ‘말(더러브렛종) 유전능력 평가’ 용역 과제로 진행되었음.

VI. 인용 문헌

1. Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Second international symposium on information theory. Paper No. 267-281.
2. Bakhtiari, J. and Kashan, N. E. J. 2009. Estimation of genetic parameters of racing performance in Iranian thoroughbred horses. *Livestock Science*. 120(1-2):151-157.
3. Chico, M. D. 1994. Genetic analysis of thoroughbred racing performance in Spain. *Ann Zootech*. 43:393-397.
4. Ekiz, B. and Kocak, O. 2007. Estimates of genetic parameters for racing times of thoroughbred horses. *Turk. J. Vet. Anim. Sci*. 31(1):1-5.
5. Meyer, K. 2007. WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *J Zhejiang Univ Sci B*. 8(11):815-821.
6. Montgomery, D. C. 2005. Design and analysis of experiments. 6th ed., Wiley., p. 94-95.
7. Moritsu, Y., Terai, A. and Tashiro, T. 1998. Relationship between sire breeding values for the rating score on turf and dirt racing tracks in thoroughbred racehorses. *J. Equine Sci*. 9:89-92.
8. Mota, M. D. S., Abrahao, A. R. and Oliveira, H. N. 2005. Genetic and environmental parameters for racing time at different distances in Brazilian thoroughbreds. *J. Anim. Breed. Genet*. 122:393-399.
9. Park, K. D., Lee, H. K., Jeon, G. J., Kong, H. S., Cho, B. W., Cho, K. H. and Son, S. K. 2008. Trends on racing speed traits in thoroughbred racehorses. *Korean J. Anim. Sci*. 50(6): 747-752.
10. Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*. 6:461-464.
11. Tukey, J. W. 1953. The problem of multiple comparisons. Unpublished notes, Princeton University.

(접수일자 : 2009. 2. 6. / 수정일자 : 2009. 4. 20. / 채택일자 : 2009. 4. 21.)