Livestock Science 135 (2011) 293–299

Contents lists available at ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci

Short communication

Principal components analysis applied to genetic evaluation of racing performance of Thoroughbred race horses in Korea

Jong-Eun Park [a,1], Jeong-Ran Lee [b,1], Seungyoon Oh [c], Jin Woo Lee [d], Hee-Seok Oh [b], Heebal Kim [a,c,*]

[a] Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Republic of Korea
[b] Department of Statistics, Seoul National University, Republic of Korea
[c] Interdisciplinary Program in Bioinformatics, Seoul National University, Republic of Korea
[d] Horse Registry, Korea Racing Authority, Republic of Korea

ARTICLE INFO

Article history:
Received 23 June 2009
Received in revised form 13 July 2010
Accepted 15 July 2010

Keywords:
Genetic evaluation
Heritability
Normalization
Principal components analysis (PCA)
Racing performance
Thoroughbred
Variable selection

ABSTRACT

Selection of proper phenotypic trait among various traits related with interesting performance plays an important role in genetic evaluation. In this study, principal components analysis (PCA) was adapted to generate a new index as a measure of racing performance of 12,279 horses. This method allows us to reduce the number of variables considered in the evaluation of the horses' racing performance, which may facilitate modeling genetic programs. The resulted racing time, earning prize and rank were selected for generating new indices as the representation of racing performance of the horses. Three indices used in this study were: 1) PCA1 generated from the modified values of racing time, earning prize and rank, 2) PCA2 generated from the modified racing time and rank, and 3) the adjusted racing time. The first principal components (PCs), elements in the eigenvector corresponding to the largest eigenvalue of PCA, of PCA1 and PCA2 explained the variance of the selected variables about 75.6% and 75.4% respectively. Linear combinations of the first PCs and adjusted variables were used as new performance indices. Those animal models were composed of significant explanatory variables selected by Akaike information criterion (AIC). Heritability and repeatability were 0.324 (±0.026) and 0.334 (±0.034) for adjusted racing time, 0.319 (±0.014) and 0.326 (±0.018) for PCA1, and 0.324 (±0.010) and 0.332 (±0.012) for PCA2 respectively. Estimated heritabilities and repeatabilities for three indices showed similar values for domestic racing records. However, models using PCA showed better fitting for data than model using racing time as a performance index. The proposed methodology is efficient to evaluate the total variance in this group of correlated traits, allowing reduction in the number of variables for genetic evaluation and construction of better fitting model.

1. Introduction

Thoroughbred horses have been bred exclusively for racing in England since Tudor times and Thoroughbred horse racing is now a worldwide sport and huge industry. About 110,000 foals of Thoroughbreds are registered each year all over the world (The Jockey Club, 2008); and in Korea, about 1000 foals of Thoroughbreds are registered each year by the report of International Federation of Horseracing Authorities (IFHA) in 2007.

Breeding goal of Korea Racing Authority is to produce noble, correct and durable racing Thoroughbred horses, which are internationally competitive through their temperament, racing ability and good movements. Seeking an improvement in the process of selection and breeding in Thoroughbreds is founded on the belief that racing performance is inherited (Ricard et al., 2000). Therefore, combining

* Corresponding author. Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea 151-742. Tel.: +82 2 880 4822; fax: +82 2 873 2271.
E-mail address: heebal@snu.ac.kr (H. Kim).

[1] The first two authors contributed equally to this study.

1871-1413/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.livsci.2010.07.014

those variables that represent the traits of racing performance and then figuring out the best model of explanatory variables that affect the result of racing performance is important to assess the potentials of development in Thoroughbred horse breed.

Traits related with racing performance can be measured in racing time, earning prize and rank (Langlois, 1980), and these are well studied in various studies (Hintz, 1980; Sobczynska, 2003; Mota et al., 2005; Bakhtiari and Kashan, 2009). However, contrary to the ordinary quantitative variables, these have relative values caused by environmental or economic conditions of each racing competition, such as the racing distance, the total amount of grand prize, yearly inflation rates, grade of racing group and etc. Therefore, we need careful normalization of these performance variables to consider fair comparison of individual racing records, and also have to construct a simple animal model of one combined performance index with these normalized variables. In this paper, we precisely normalized the records of racing time, ranking and earning prize with respect to the historical and relative conditions and then applied principal components analysis (PCA) to make a new performance index. PCA is the simplest of the true eigenvector-based multivariate analyses. Its operation can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data. The resulting indices contain more information that characterizes the racing ability than just a single-trait variable.

To evaluate the genetic model of racing ability, we applied estimation of variance components (VCs) for the random effects. Here, variable selection, normalization and PCA procedures were aimed at minimization of prediction error (PE) defined by the differences between estimated and true values of VCs. The selection of the influential factors which consist of environmental effect was necessary to construct accurate animal model. In this study, we constructed a reliable model of the best-fitted to the observed racing records data.

The aim of our study was to propose more suitable indices for racing performance of Thoroughbred race horses using PCA methodology and investigate the effects that influence their performance. The obtained estimates of genetic parameters and the novel procedure proposed in our study would be used for designing a breeding program for Korean Thoroughbred horses.

2. Materials and methods

2.1. Racing performance and pedigree data

All racehorses registered in the Korea Racing Authority (KRA) have been accommodated and trained within the confines of KRA's facilities for racing. Total 220,866 racing data of 12,279 racehorses recorded from the competitions held in Seoul and Busan horse racing park provided by KRA were used in this study. All of the races took place on dirt tracks. Racing time, rank at finish and earning prize were recorded for all horses participating in the race. The records of horses that could not complete the race were deleted from the dataset. Description of the number of animals according to race distance and sex are illustrated in Table 1. The average number of animals/race for Seoul and Busan racing park was 9.52 and 10.55 respectively. The average number of races/animal for Seoul and Busan racing park was 18.97 and 12.74 respectively. The collected data covered in the period of 19 years, from 1990 to 2008. All the used data belong to domestic races though there were diversities in the country of origin of horses. Pedigree information of horses covered 9 generation and the number of animals in relationship matrix was 66,644.

2.2. Preprocessing and variable normalization

The three candidates for performance indices needed more careful adjustments. Racing time takes higher values when the total distance of the race is longer, so we adjusted the time records by subtracting the difference between the fitted values of each distance in linear regression model:

$$\text{Fitted me an of racing time record by distance} = -6.778 + 0.071 \times \text{racing distance}.$$

Table 1
Number of horses (No. of horses) and records (racing time, rank and earning) for different distances.

Distance (m)	No. of horses				No. of records (racing time, rank and earning)			
	Total	Female	Male	Gelding	Total	Female	Male	Gelding
Total (1)	12,279	6400	3668	2800	220,866	112,288	55,064	53,514
1000 (2)	10,945	5828	3246	2261	46,244	26,595	11,770	7879
1200 (2)	10,642	5578	3093	2263	52,714	28,654	13,463	10,597
1300 (3)	932	433	308	209	2135	1038	651	446
1400 (2)	8877	4534	2411	2094	48,065	25,579	11,368	11,118
1600 (4)	933	443	369	128	3932	1864	1496	572
1700 (3)	5331	2569	1354	1455	20,297	9616	4976	5705
1800 (2)	4869	2230	1329	1368	23,427	10,573	5813	7041
1900 (3)	2649	1152	625	891	11,480	4394	2641	4445
2000 (2)	2078	844	531	723	11,439	3633	2631	5175
2200 (2)	143	54	29	60	350	124	80	146
2300 (3)	368	109	87	173	783	218	175	390

(1) One horse has run several distances. And some male horses were gelded, a horse with an ID can have two different sexes, male and gelding.
(2) Held in both hippodroms.
(3) Held in Seoul park only.
(4) Held in Busan park only.