Journal of Animal Science and Technology 52(6) 459~466, 2010
DOI:10.5187/JAST.2010.52.6.459

Prediction Models for Racing Performance of Domestic Progeny of Thoroughbreds

Jeong-Ran Lee[1], Jin Woo Lee[2], Heebal Kim[3] and Hee-Seok Oh[1]*

[1]Department of Statistics, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-747, Korea.

[2]Horse Registry, Korea Racing Authority, 685 Juam-dong, Gwacheon-si, Gyeonggi-do 427-711, Korea.

[3]Department of Agricultural Biotechnology, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-747, Korea.

ABSTRACT

In this study, we suggest an objective standard in selection of candidate horse mates. Korea Racing Authority provided racing records and pedigree information of 44 sires and 954 dams. The datasets were used to predict Racing Indices represented by the averages of earnings earned by offspring for each dam and sire that indicate the racing performance of its domestic progeny. Proportion of wins and second places to the number of taken races and the mean of distances for the won races of a sire were significant factors in linear model with minimum prediction errors. For dam, those factors were the average of earned money per race, number of outstanding broodmares in pedigree, and the comparable index which indicates the relative affinity with its mate. We can use the resultant model for a horse mate by choosing one of the candidates with the largest predicted value for hypothetical offspring.

(**Key words :** Cross-validation, Horse breeding, Linear model, Statistical learning, Thoroughbred)

INTRODUCTION

The male parent of a horse, a stallion, is commonly known as the sire and the female parent, the mare, is called the dam or broodmare. Both are genetically important, as each parent provides half of the genetic makeup of the ensuing offspring. They often are chosen in hopes of passing down their outstanding physical or athletic attributes, or desirable ancestry.

The Thoroughbred is a horse breed best known for its use in horse racing because it has agility, speed and spirit. Thoroughbreds 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). In Korea, about 1,000 foals of Thoroughbreds are registered each year by the report of International Federation of Horseracing Authorities (2007). Korea Racing Authority (KRA), the nation's authorized horse racing institute, is entitled to restore the national identity of horse racing. Breeding goal of KRA is to produce noble, correct and durable racing Thoroughbreds, which are internationally competitive through their temperament, racing ability and good movements. Thus, related research is in demand by the industry.

There are many previous studies which estimate genetic parameters and evaluate fixed effects relative to racing performance (Lee et al., 1995; Park and Lee, 1999; Bakhtiari and Kashan, 2009). However, since those literatures focused on estimating the breeding values using animal mixed models, the proper use of statistical analysis has not been applied extensively to identifying the linkage of offspring with their parents for specific characteristics in individuals of differing phenotypes that most affect the progeny's racing performance. The goal of this study is to select significant factors in racing and pedigree records of each sire and dam, and to predict the offspring's racing performance in the statistical learning framework.

MATERIALS AND METHODS

1. Data structures and terminologies

There are observations of 44 sires and 954 dams with domestic and foreign records, but those for whose offspring have domestic racing records only. This data, covered in the

* Corresponding author : Hee-Seok Oh, Department of Statistics, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-747, Korea. Tel: 82-2-880-9242, Fax: 82-2-883-6144, E-mail: heeseok@stats.snu.ac.kr

period of 20 years from 1990 to 2009, is provided by KRA for an analysis. We focus on finding out significant factors in the records of each sire and dam that can be used to predict the quality of its hypothetical offspring. A component of the response variable is the offspring's performance presented in terms of Racing Indices (RI). Each offspring's RI is calculated by the ratio of its mean earnings per race over that of its contemporary group. The contemporary group is specified based on its element offspring's racing information in Korea defined by a complex scheme of KRA's private customary rules, which consists of age (2, 3, 4 or over 5), country of origin (domestic or foreign), hippodrome (Seoul or Busan), and the year of the race is held. The value of RI greater than one indicates that the offspring has earned more the mean earnings per race than the average of that earned by its contemporary group. We obtain the response variable as a form of Average Performance Indices (API) per individual sire or dam by taking averages of RI from its entire offspring.

Short descriptions of all candidate explanatory variables are presented in Table 1 for sire records and Table 2 for dam with the abbreviated variable names to be used in the rest of this article. The input variables can be explained based on what kind of traits they express in each sire and dam to be assessed. Because phenotypes of an individual sire or dam are determined by the expression of an organism's genes as well as the influence of environmental factors and the interactions between the two, each explanatory variable can be interpreted in a relationship with genetic information or aptitude in racing. For example, CI in Table 1 can be interpreted as a relative affinity with its mate of a sire or dam. If the offspring out of progeny produced by other sires with dams who are mated to the chosen sire performed well, it is possible that its own offspring result in good returns. This indicates that the relative affinity of a sire or dam with its mates can be comparable to the performance of its own offspring. Therefore, we additionally displayed possible traits related to each candidate explanatory variables in those tables.

More detailed descriptions for some candidates of explanatory variables presented in Table 1 are as follows. Average Earning Index (AEI) is computed by taking the yearly average of a proportion of the mean of earned money at each year per offspring to that of entire runners at the year. The meaning of 'outstanding' in the description of DYIELD indicates that a dam yielded offspring of grand or special race winners under the condition of having 70% of WIN_R over at least three taken races. The types of races such as grand, special or stakes are determined by the magnitude of total prize money, which are designated by KRA or other sponsors.

Additional explanations for some candidates of input variables displayed in Table 2 are as follows. The seven categories in CLASS are coded as 1 for winners of the

Table 1. Data structure for sire records

Type	Related trait	Variable name	Descriptions
Response variable	Offspring performance	API	Average of RI for all offspring by each sire.
Candidate explanatory variables	Relative affinity	CI	Comparable Index: API for offspring out of progeny by the chosen sire and mated dams when bred to all other sires.
	Excellence in racing	WIN_R	Proportion of wins to the number of taken races.
		DWIN_R	Proportion of wins and second places to the number of taken races.
		AVG_PRZ	Average of earned money per each taken races.
	Aptitude in racing	AWD	Mean of distances for the won races.
		ROAD	Track aptitude coded as 1 for turf and 2 for dirt.
	Prematurity	SEC_PRZ	Average of earned money per each taken race at two years old.
	Quality in pedigree	GRD_AEI	AEI of grandsire.
		BMS_AEI	AEI of broodmare sire.
		DYIELD	Number of outstanding broodmares in pedigree for four generations.
		SW_PED	Number of stakes winners in grand-dam or siblings of the sire.

Table 2. Data structure for dam records

Type	Related trait	Variable name	Descriptions
Response variable	Offspring performance	API	Average of RI for all offspring by each dam.
Candidate explanatory variables	Relative affinity	CI	Comparable Index: API for offspring out of progeny by the chosen dam and mated sires when bred to all other dams.
	Excellence in racing	CLASS	Coded index of racing ability categorized by type of races.
		AVG_PRZ	Average of earned money per each taken races.
	Aptitude in racing	AWD	Mean of distances for the won races.
	Genetic aptitude	INBREED	Coefficient of inbreeding.
	Prematurity	SEC_PERF	Classification of racing performance at two years old.
	Quality in pedigree	BMS_PRC	Average of mating fees paid to broodmare sire.
		BMS_AWD	AWD of broodmare sire.
		BMS_CLS	CLASS of broodmare sire.
		DYIELD	Number of outstanding broodmares in pedigree for four generations.
		SW_PED	Number of stakes winners in grand-dam or siblings of the dam.

grand races, 2 for second or third places of the grand races, 3 for stakes winners or fourth places of the grand races, 4 for ones have earnings in stakes races except winners, 5 for winners of races except grand and stakes, 6 for ones had taken races at least once, and 7 for those never had taken races, respectively. Inbreeding is a genetic term that refers to reproduction as a result of the sexual intercourse of two animals which are genetically related to each other. If the relationship is a close one or it is practiced repeatedly, inbreeding can increase the chances of offspring being affected by recessive or deleterious traits. The percentage of chances for two alleles to be identical by descent is called inbreeding coefficient. For more details in calculation of inbreeding coefficient, refer Wright (1922). The three categories of SEC_PERF are coded as 1 for winners, 2 for ones had taken races at least once, and 3 for those never had taken races, respectively. Because prize money have the unit of 1,000 Korean Won (KRW), other currencies used in the representation of mating prices for BMS_PRC or AEI are converted to KRW based on current exchange rates at the corresponding time.

2. Preprocessing

There were missing values of 717 (75.16%) observations for BMS_PRC, 204 (21.38%) for BMS_AWD, 58 (6.08%) for AVG_PRZ, and 243 (25.47%) for AWD among all candidate explanatory variables in dam data. Because the missing proportion of BMS_PRC was huge, the factor itself was excluded in the analysis. Factors with missing values of over twenty percents, i.e., BMS_AWD and AWD were also not included in models because they were not significant ($P > 0.05$) in the preliminary testing on datasets with no missing values. Outliers positioned over 99% of entire data were also eliminated. Therefore, we conducted our analysis on 873 records of dam resulted from the removed observations with those missing values in variables AVG_PRZ, and outliers. There were no missing values in records of sires. We standardized all factors by converting to z-scores of variables (which are derived by subtracting the mean from an individual raw score and then dividing the difference by the standard deviation) because the scales of them were all different. For those variables related to average earnings, we took the natural logarithm (ln) after increasing them by one to normalize because there were too many zeros and extreme values.

3. Framework of statistical learning

The science of learning plays a key role in the fields of statistics, data mining and artificial intelligence, intersecting with areas of engineering and other disciplines. It is about learning from data. In a typical scenario, we have an outcome measurement, usually quantitative or categorical, that we wish to predict based on a set of inputs, called features. We have a training set of data, in which we observe the outcome and feature. Using this data we build a prediction model, or learner, which will enable us to predict the outcome for new unseen objects. A good learner is one that accurately predicts such an outcome, and the performance of a learner can be validated in a set of test data. Various aspects and examples of statistical learning can be found in Hastie et al. (2009).

We denote a sample of input-output pairs ($\boldsymbol{X}_i$, Y_i), i=1, ⋯, n, where $\boldsymbol{X}_i = (X_{i1},\cdots,X_{id})$ denotes a d-dimensional covariate vector and Y_i denotes a continuous response. Each method we will use has complex parameters, and those parameters are chosen to minimize an estimate of prediction error based on tenfold cross-validation (CV). Tenfold CV works by dividing the training data randomly into ten equal parts. Our learning methods are fit to nine-tenths of the data, and prediction errors are computed on the remaining one-tenth. This is done in turn for each one-tenth of the data, and the ten prediction error estimates are averaged. The measure of prediction error to minimize is the predicted mean squares error (PMSE) defined as follows:

$$PMSE_t = \frac{1}{n_t} \sum_{i=1}^{n_t} (Y_i, \hat{Y}_i^{(-n_t)})^2$$

where Y_i is the ith observation of the test set (size of n_t), which indicates the true value of the response, and $\hat{Y}_i^{(-n_t)}$ is the predicted value for test set covariates obtained in the model from training set. Here, t equals ten if tenfold CV is used. For a simple notation of the formula, we denote CV_k for the mean of $PMSE_t$, $t = 1$, ⋯, k. Because there are 44 observations of sires and 873 preprocessed records of dams, we partition each datasets into 40 vs. 4 and 776 vs. 97, respectively, to make equally spaced folds for the convenience of computation. The consequence is that we used eleven-fold CV for sire data and nine-fold CV for dam data. We can capture whether the model is stable, and detect the most accurate model as well by changing partitions in the nine-fold or eleven-fold CV.

The three different prediction models we present in this article are obtained by applying the linear regression model, tree-based method, and an ensemble method called bagging to datasets. Validations of proposed models are represented in terms of PMSE. Application and evaluation of the proposed methods are implemented by R, a free software environment for statistical computing and graphics obtained from the Comprehensive R Archive Network (CRAN).

4. Linear model and variable selection

Linear regression model assumes that the regression function $E(Y \mid \boldsymbol{X})$ is linear in the parameter β_js (called coefficients, which is unknown constants to be estimated from data) such as:

$$Y = E(Y \mid \boldsymbol{X}) + \varepsilon = \beta_o + \Sigma^{d}_{j=1} X_j \beta_j + \varepsilon,$$

where, the error ε is assumed to be a Gaussian random variable with mean zero and variance σ^2. It is simple, and often provides an adequate and interpretable description of how the inputs affect the output. For example, if we get an estimate $\hat{\beta}_j$ of β_j as a constant 2, then we can interpret that the one unit increment in X_j results in 2 unit increment in the expected value of the response. For prediction purposes, it can sometimes outperform complicated nonlinear models, especially in situations with small numbers of training cases or sparse data.

We conduct variable selection to obtain a stable predictor for any changes in datasets by reducing model complexity. We first identify significant factors among the categorical variables with levels more than two by a single factor ANOVA. And then, we apply a well-known stepwise variable selection procedure for the linear model in combination with Akaike Information Criterion (AIC; Akaike, 1973) for fifty times, and count the numbers of being selected for each variable. Among the chosen variables, we make combinations of those candidate explanatory variables to construct possible linear models, and find the best combination that provides the smallest PMSE.

5. Tree-based method

Regression tree is a nonparametric regression method that partitions the feature space of $\boldsymbol{X}$ into a set of rectangles and takes the average of the response values in each rectangle as an estimate of the regression function in that region. It is conceptually simple and powerful. We use the procedure called Classification and Regression Trees (CART), which is a popular method for tree-based regression. Its construction is based on recursive binary partitioning of the feature space with additional growing and pruning stages. A key advantage of the recursive binary tree is its interpretability, because the partition of feature space is fully described by a single tree and the representation works in the same way with more than two inputs.

For each variable X_j, define

$$R_1(j, s) = \{\boldsymbol{X} : X_j \leq s\} \text{ and } R_2(j, s) = \{\boldsymbol{X} : X_j > s\},$$

which are the two regions split by the variable j at the point s. Also, define

$$\overline{Y}_1(j, s) = \text{average } \{Y_i \mid \boldsymbol{X}_i \in R_1(j, s)\} \text{ and}$$
$$\overline{Y}_2(j, s) = \text{average } \{Y_i \mid \boldsymbol{X}_i \in R_2(j, s)\}.$$

In growing stages, we find (j, s) that minimizes the residual sum of squares of the best constant fit such that:

$$\Sigma_{i:\boldsymbol{X}_i \in R_1(j, s)} \{Y_i - \overline{Y}_1(j, s)\}^2 + \Sigma_{i:\boldsymbol{X}_i \in R_2(j, s)} \{Y_i - \overline{Y}_2(j, s)\}^2.$$

And then repeat the splitting process by partitioning the feature space into the two resulting regions until some stopping rule is applied. Too large trees with too many terminal nodes may overfit the data, while too small trees may not capture the important structure of the regression function. Therefore, we chop off the last grown node first and so on, and then obtain a sub-tree by pruning a tree given from the growing stage. In contrast to the previous linear model, there is no need to select significant variables in advance. For more details in selection of the best sub-tree that minimizes the cost complexity criterion, see Breiman et al. (1984) and Ripley (1996).

6. Ensemble method

Ensemble is a generic term for methods of constructing many learners and combining them to make a highly accurate learner. Here, we use an example called bagging (i.e., bootstrap aggregating) introduced by Breiman (1996). Empirically, ensemble methods perform better than the best single learner, particularly when the learner is unstable. Because the bootstrap is a way of assessing the accuracy of estimation or prediction, bagging can serve as a method of improving prediction.

Let $f(\cdot, \boldsymbol{S})$ denote a regression estimate based on a sample $\boldsymbol{S}$. We draw bootstrap samples (by random sampling with replacement) $\boldsymbol{S}_{(b)}$, $b = 1, \cdots, B$, from the training sample. The bagging estimate is computed by

$$f_{bag}(\boldsymbol{X}) = \frac{1}{B} \sum_{b=1}^{B} f(\boldsymbol{X}, \boldsymbol{S}_{(b)}).$$

Here, we choose the base learner as the tree-based estimate. It can be seen in the formula of $f_{bag}(\boldsymbol{X})$ that the interpretation of the model is not simple though the performance might be better than that of a single learner.

RESULTS & DISCUSSION

1. Sire assessment model

Among fifty times of the stepwise variable selection procedure, only DWIN_R (96%), AVG_PRZ (38%), AWD (38%), and SEC_PRZ (30%) are significant ($P < 0.05$) in linear models for sire records. Though the importance of the variable DWIN_R is obvious, we further need to compare the model performance for all combinations of the rest three variables. Because the selected percentage of variables AVG_PRZ and AWD is the same, we consider the four linear models of variable combinations as displayed in Table 3. The model with DWIN_R and AWD has performed best with the smallest CV_{11} (= 0.944) among linear models. The resultant model containing only those variables with the minimum $PMSE_3$ (= 0.225) is:

$$\text{API} = 0.231 \times (-7.302 \times 10^{-18} + 0.327 \times \frac{\text{DWIN\_R} - 0.377}{0.124}$$
$$- 0.249 \times \frac{\text{AWD} - 1534.6}{438.478}) + 0.895.$$

The negative coefficient for the z-score of AWD indicates that those sires that have shorter average winning distance produce better performed offspring.

The summary statistics of $PMSE_i$s result from tree-based and bagging models for sire records by eleven-fold CV are also presented in Table 3. The models with bagging result in smaller CV_{11} than that of CART because $PMSE_i$s of CART are obtained by single learners which based on trees, though both of them performed worse than linear models. The resultant model with the minimum $PMSE_9$ (=0.974) obtained from CART is given in Fig. 1, for example. It can be interpreted that the proportion of wins to the number of taken races splits the inputs first, and the average winning distance does in consecutive order. This also indicates the

Table 3. Performances of proposed prediction models by eleven-fold CV in sire records

Summary statistics of $PMSE_1, \cdots, PMSE_{11}$	Linear model				CART	Bagging
	DWIN_R, AWD, AVG_PRZ, SEC_PRZ	DWIN_R, AVG_PRZ, AWD	DWIN_R, AVG_PRZ	DWIN_R, AWD		
Minimum	0.231	0.371	0.318	0.225	0.974	0.351
First quantile	0.546	0.549	0.625	0.054	1.242	0.486
Median	0.899	0.913	0.793	0.951	1.543	0.982
Mean (= CV_{11})	0.982	0.993	0.963	0.944	1.731	1.074
Third quantile	1.291	1.331	1.162	1.213	2.152	1.524
Maximum	2.140	2.349	2.389	2.394	2.947	2.108

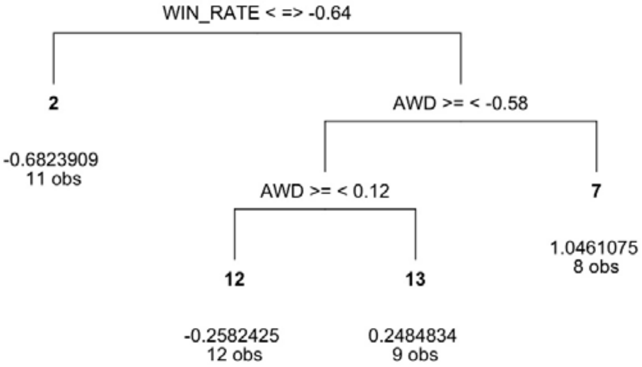

Fig. 1. An example of tree-based prediction model for sire records with minimum PMSE. The split values correspond to z-scores of each variable.

better accuracy of linear model in the above example, because both used the same explanatory variables. The resultant minimum PMSE model of bagging is obtained by ensemble trees of twenty five bootstrapped nodes, so the interpretation is very difficult. It can be only mentioned that though there are some variations, the majority of nodes in the best bagging model consist of CI, AVG_PRZ, AWD, DWIN_R, and SEC_PRZ, which are similar to those selected in the linear models.

2. Dam assessment model

The categorical variables with levels more than two in dam dataset are CLASS, SEC_PERF and BMS_CLS. We first conduct a single factor ANOVA on these factors and find that they are not significant ($P > 0.1$). Among the fifty stepwise procedures, only AVG_PRZ (100%), DYIELD (94 %), and CI (100%) are selected as significant ($P < 0.05$) for linear model of dam records. The nine-fold CV results are displayed in Table 4. The resultant linear model, for example, with the minimum $PMSE_1$ (=0.569) is:

$$\text{API} = 0.763\times(-0.515+0.057\times\ln(\text{AVG\_PRZ}+1) + 0.058\times\frac{\text{DYIELD}-2.966}{1.803} + 0.162\times\frac{\text{CI}-1.013}{0.155}) + 1.002.$$

Table 4. Performances of proposed prediction models by nine-fold CV in dam records

Summary statistics of $PMSE_1, \cdots, PMSE_9$	Linear model	CART	Bagging
	AVG_PRZ, CI, DYIELD		
Minimum	0.569	0.630	0.630
First quantile	0.642	0.681	0.656
Median	0.924	1.086	1.003
Mean (= CV_9)	0.980	1.044	1.001
Third quantile	1.159	1.183	1.163
Maximum	1.918	1.978	1.922

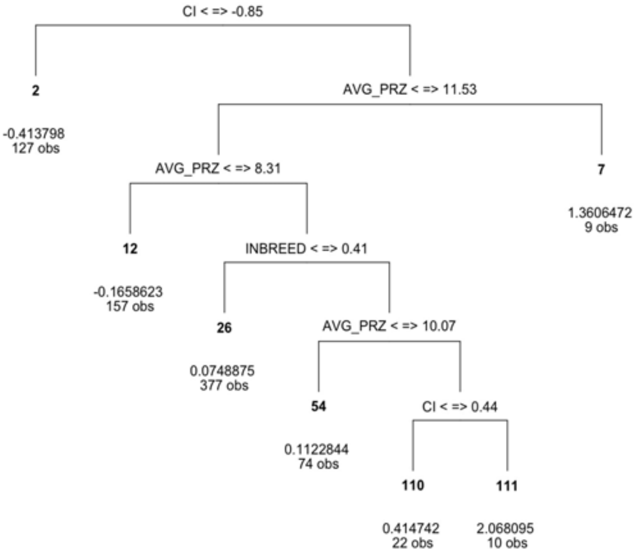

Fig. 2. An example of tree-based prediction model for dam records with minimum PMSE. The split values correspond to z-scores of each variable.

And an example of CART with the minimum PMSE is in Fig. 2, which is more complex than that of sires. Its interpretation can be done similar to the case of sires. The linear prediction model performed the best for dam data as in Table 4.

3. Implications

A stallion with a proven competition record is one criterion for being a suitable sire. The stallion should be chosen to complement the mare, with the goal of producing a progeny that has the best qualities of both animals, yet avoids having the weaker qualities of either parent. Some breeders consider the quality of the sire to be more important than the quality of the dam, and other breeders maintain that the mare is the most important parent. Because stallions can produce far more offspring than mares, a single stallion can have a greater overall impact on a given phenotype, or breed. However, the mare may have a greater influence on an individual progeny because its physical characteristics influence the developing foal in the womb and the foal also learns habits from its dam when young. Foals may also learn the language of intimidation and submission from their dam, and this imprinting may affect the offspring's status and rank within the herd. Many times, a mature horse will achieve status in a herd similar to that of its dam; the offspring of dominant mares become dominant themselves.

Here we cannot tell whether which parent (sire or dam) has more impact on its offspring than the other, the results from the proposed methods provide most important factors in the assessment of itself. For sire records, the individual excellence or aptitude in racing of each sire itself are most important factors in producing dominant offspring. In contrast, the affinity and genetic related factors influence to the offspring's racing performance for dam records, indicates that the quality in pedigree and mating plans need to be considered as a prior factors in the assessment of a dam. Furthermore, we can explicitly predict the expected value of a hypothetical offspring's racing performance index from contributions of only selected variables when we use linear models. If the proposed methods are combined with an analysis of breeding values, more effective models can be obtained because it is based on racing records that might complement the limitation of information that only depends on the RI commonly used in nicking systems to present horses' racing performance. This remains as our further research.

ACKNOWLEDGEMENTS

This work was supported in resources and funded by Korea Racing Authority for the project of Horse (Thoroughbred) Genetic Evaluation.

REFERENCES

Akaike, H. 1973. Information theory and the maximum likelihood principle. Page 267 in Second International Symposium on Information Theory. B. Petrov and F. Csaki, ed. Akademiai Kiado, Budapest.

Bakhtiari, J. and Kashan, N. E. J. 2009. Estimation of genetic parameters of racing performance in Iranian Thoroughbred horses. Livestock Science 120:151-157.

Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. 1984. Classification and Regression Trees. Chapman & Hall, New York.

Breiman, L. 1996. Bagging predictors. Mach. Learn. 26:123-140.

CRAN. The Comprehensive R Archive Network: http://cran.r-project.org/.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, New York.

International Federation of Horseracing Authorities. 2007. IFHA 2007 Annual Report:
http://www.horseracingintfed.com/.

Lee, K. J., Park, K. D., Kang, M. G., Kim, D. R. and Moon, Y. Y. 1995. Estimation of genetic parameters for racing performance of Thoroughbred horses. Kor. J. Anim. Sci. 37:11-18.

Park, K. D. and Lee, K. J. 1999. Genetic evaluation of Thoroughbred racehorses in Korea. Kor. J. Anim. Sci. 41:135-140.

Ripley, B. D. 1996. Pattern Recognition and Neural Networks. Cambridge University Press, New York.

The Jockey Club. 2008. Thoroughbred Racing and Breeding Worldwide:
http://www.jockeyclub.com/factbook.asp?section=17.

Wright, S. 1922. Coefficients of inbreeding and relationship. Am. Nat. 56:330-338.

(Received Oct. 4, 2010; Revised Nov. 24, 2010; Accepted Dec. 1, 2010)