The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1133
28(6), 1133–1146

Analysis of Horse Races: Prediction of Winning Horses in Horse Races Using Statistical Models

Hyemin Choe[a] · Nayoung Hwang[a] · Chankyoung Hwang[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received September 14, 2015; Revised October 13, 2015; Accepted October 19, 2015)

Abstract

The Horse race industry has the largest proportion of the domestic legal gambling industry. However, there is limited statistical analysis on horse races versus other sports. We propose prediction models for winning horses in horse races using data mining techniques such as logistic regression, linear regression, and random forest. Horse races data are from the Korea Racing Authority and we use horse racing reports, information of racehorses, jockeys, and horse trainers. We consider two models based on ranks and time records. The analysis results show that prediction of ranks is affected by information on racehorses, number of wins of racehorses and jockeys. We place wagers for the last month of races based on our prediction models that produce serious profits.

Keywords: horse race, linear regression, stepwise regression, random forest, logistic regression, important variables

1. 서론

국내 합법 사행산업은 기존에 있던 경마, 경륜, 복권, 카지노(외국인 대상)에서, 2000년 이후 내국인 카지노, 스포츠 토토, 경정, 온라인 복권 등으로 확대되어 왔다. 2014년 매출액을 기준으로 보면 경마가 38.4% (The National Gambling Control Commission, 2015)로 사행산업 중 가장 큰 비중을 차지하고 있다. 또한 2014년 국내 주요 스포츠 관람객 수를 살펴보면 야구는 675만명, 축구는 186만명 (Statistics Korea e-National indicators, 2015)으로 나타났고, 경주 스포츠인 경마는 1,529만명, 경륜은 529만명 (The National Gambling Control Commission, 2015)으로 관람 스포츠 중 경마의 관람객이 가장 많다. 과거에는 경마에 대한 부정적 인식이 강했으나 최근에는 경마가 놀이문화로서 자리 잡고 있는 추세이다. 그 예로 경마의 도박 중독률이 2012년 60.3%에서 2014년 49.1%로 2년 만에 11.2% 감소했으며 (The National Gambling Control Commission, 2014), 고액 배팅인 10만원권 구매 비율이 2004년 6.6%에서 현재 3.1%로 절반 이상 줄었고, 같은 기간 3천원 이하 소액 구매 비율이 20.4%에서 30.8%로 1.5배 늘었다는 연구 결과가 있다 (The Korea Racing Authority, 2014).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education, Science and Technology (No. NRF-2013R1A1A2 012817).

[1]Corresponding author: Department of Statistics, Ewha Womans University, Seoul 03760, Korea.
E-mail: josong@ewha.ac.kr

현재 한국 마사회에서는 해당 경기에 대한 출전표, 경주마, 기수, 조교 정보 등 다양한 데이터를 사전에 제공하고 있어 데이터에 접근이 용이하다. 그러나 관련 데이터로 통계적 예측 모형을 활용한 분석이 타 스포츠 종목에 비하여 이루어진 사례가 적다 (Yoo와 Park, 2000). 본 연구에서는 데이터 마이닝 기법을 이용하여 경마 순위 예측 모형을 제안하고자 한다.

현재 국내 경마 경기는 매주 이틀씩 서울, 부산, 제주에서 각각 개최되고 있으며, 이 중 매출액의 55%를 차지하고 있는 서울 지역 경기를 분석 대상으로 하였다. 분석에 이용한 데이터는 한국마사회 홈페이지 자료실에서 제공하는 2014년 1월부터 2015년 4월까지의 경마 성적표, 경주마, 기수, 조교사 정보를 이용하였다. 경마 순위 예측모형을 위하여 본 논문에서는 두 가지 모형을 고려하였다. 첫 번째는 경기 결과의 순위를 기반으로 한 모형으로, 다양한 분류분석방법을 이용하여 예측 모형을 적합하였다. 두 번째는 기록을 기반으로 한 모형으로, 우선 경기 기록을 예측한 후 기록에 따른 순위를 부여하는 방법을 이용하였다. 경마 특성상 배팅이 주목적이므로 우리는 배팅 방식에 따라 원하는 순위까지 예측하였을 때의 예측 정확성(예측률)을 계산하여 비교하고자 한다. 분석에는 선형 회귀 모형, 로지스틱 회귀 모형 (McCullagh와 Nelder, 1989; Hastie와 Pregibon, 1992), 랜덤 포레스트 모형 (Breiman, 2001)을 이용하였다. 분석은 R (R Development Core Team, 2010)을 이용하여 이루어졌으며, R에 내장된 다양한 함수와 패키지를 이용하여 주요변수를 선택하고 모형을 적합하였다.

본 논문의 순서는 다음과 같다. 2장에서는 자료 수집 방법과 분석에 사용된 변수에 대하여 설명하고, 3장에서는 분석 방법과 분석 결과를 비교하여 최적 예측 모형을 제시하고 최근 자료에 모형을 적용한 결과를 비교한다. 4장에서는 본 연구의 내용을 요약하고 결론을 내리고자 한다.

2. 분석자료 설명

2.1. 용어 정의

한국마사회법에 따르면 경마란 기수가 기승한 말의 경주에 대하여 승마투표권을 발매하고, 승마투표 적중자에게 환급금을 지급하는 행위를 말한다. 본 연구에서는 경주마의 순위 예측 모형의 평가 지표로 경마 배팅 방식을 이용하였다. 경마 배팅 방식으로는 단승식, 연승식, 복연승식, 복승식, 쌍승식, 삼복승식 등이 있으며, 각 배팅방식은 다음과 같다.

- 단승식: 1등으로 도착할 말 1두를 적중시키는 방식.
- 연승식: 1~3등 안에 들어올 말 1두를 적중시키는 방식 (경주에 출주하는 마필이 7두 이하일 때에는 2등 안에 들어올 말 1두를 적중시키는 방식).
- 복연승식: 1~3등 안에 들어올 말 2두를 순서에 상관없이 적중시키는 방식.
- 복승식: 1등과 2등으로 들어올 말 2두를 순서에 상관없이 적중시키는 방식.
- 쌍승식: 1등과 2등으로 들어올 말 2두를 순서대로 적중시키는 방식.
- 삼복승식: 1등, 2등 및 3등으로 들어올 말 3두를 순서에 상관없이 적중시키는 방식.

2.2. 자료수집

본 연구에 사용된 자료는 2014년 1월 4일부터 2015년 5월 31일까지 서울경마공원에서 실시된 경마 경주 자료로, 한국 마사회(www.kra.co.kr)의 공공데이터포털을 이용하여 자료를 수집하였다. 총 18,062개의 자료 중 2014년 1월부터 2015년 4월까지의 16,821개의 자료(총 1,474개의 경기)를 train data로, 2015년 5월에 해당하는 1,241개의 자료(총 109개의 경기)를 test data로 설정하였다. 분석을

Table 2.1. The number of games by distance

거리(m)	1000	1100	1200	1300	1400	1700	1800	1900	2000	2300
자료 수	1812	283	4520	3497	3257	753	1939	442	297	21
경기 수	166	29	402	312	273	68	160	37	25	2

Table 2.2. A race track's condition according to humidity

주로 습도	1∼5%	6∼9%	10∼14%	15∼19%	20% 이상
주로 상태	건조	양호	다습	포화	불량

위하여 각 경주일마다의 경마 성적표와 경주마 정보, 기수 정보, 조교사 정보를 사용하였다. 먼저 경마 성적표를 통하여 경주 거리, 날씨, 주로 상태, 주로 습도, 경주마의 순위, 경주마 번호, 산지, 성별, 나이, 부담중량, 기수, 조교사, 마주, 마체중, 경주 기록 자료를 수집하였고 이에 해당하는 기수 정보와 조교사 정보, 경주마 정보를 수집하였다. 각 변수에 대한 자세한 설명은 다음 절에서 하고자 한다.

2.3. 변수 설명

본 연구의 목적은 경주마에 대한 정보와 기수 정보, 조교사 정보를 가지고 우승마를 예측하는 것이다. 반응변수는 경주마의 순위와 기록 두 가지로 설정하였다.

자료 수집 시 다음과 같은 결측치가 발생하여 제거하였다. 먼저 경기 도중 실격이나 출전중지 등의 사유로 순위가 존재하지 않는 자료 308건과 조교정보가 존재하지 않는 자료 6건을 제거하였다. 다음으로 경주마의 등급이 정해지지 않은 말은 보통 비슷한 등급의 말로 경기가 구성되기 때문에 같은 경기에 출전하는 말의 등급의 최빈값으로 결측치를 대체하였다. 또한 1000m, 1300m, 1700m, 1800m에서 이상치가 존재하여 해당 자료를 제거한 후 분석하였다. 분석에 이용한 설명변수들은 다음과 같다.

2.3.1. 거리 경주마는 벌어들인 상금에 따라 뛸 수 있는 경주가 구분되며, 각 마필의 거리별 적성을 감안하여 3∼9개의 경주거리에 선택 출주할 수 있다. 운영거리는 1000m, 1100m, 1200m, 1300m, 1400m, 1700m, 1800m, 1900m, 2000m, 2300m 총 10가지가 있으며 각 경기에 해당하는 자료의 수와 경기의 수는 Table 2.1과 같다.

우리는 전체 자료를 이용하여 바로 순위를 예측하는 방법과 거리 별로 기록을 예측하는 회귀모형을 적합하여 순위를 예측하는 방법 두 가지를 모두 이용할 것이다.

2.3.2. 주로상태, 주로습도 경주로 상태는 경주 결과에 변수로 작용한다고 할 수 있다. 말에 따라 경주로 상태에 민감하게 작용하는 말이 있어 주로 상태가 불량일 경우 평상시와는 경주성적이 많이 달라질 수 있고, 포화 상태의 경주로에서는 빠른 주파기록이 탄생하는 경우가 많다. 또한 Table 2.2와 같이 주로의 습도에 따라서 주로 상태가 결정되기 때문에 데이터 분석모형에는 주로 상태 변수를 제외하고 주로 습도 변수만을 포함시켰다.

2.3.3. 경주마 관련 정보

- 마번

  마번은 경주마에게 부여되는 출발번호로 번호가 작은 말이 안쪽에서 출발하며 크게 두 가지 추첨방식(추첨기/컴퓨터)에 의해 랜덤하게 결정된다.

Table 2.3. The number of observations by horse's age

연령	2세	3세	4세	5세	6세	7세	8세	9세	합계
자료 수	1545	6814	5150	2405	707	161	38	1	16821

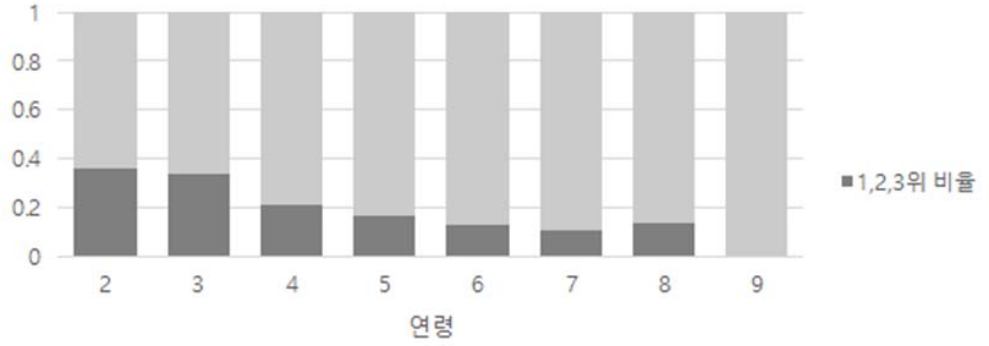

Figure 2.1. Rates of the first, second and third rank according to horse's ages.

1000m∼1400m 경주의 경주전개는 직선-곡선-직선, 1700m∼2300m 경주의 경주전개는 직선-곡선-직선-곡선-직선으로 이루어져 있으므로 바깥 라인보다 안쪽 라인에서 출발하는 경주마의 기록이 더 좋을 것으로 예상된다.

- 성별

  경주마의 성별로 암말, 수말, 거세마로 구분된다. 자료에서 암말은 7325마리, 수말은 6016마리, 거세마는 3480마리이다.

- 연령

  경주마의 연령은 2세부터 9세까지 존재하며, 각 연령에 해당하는 자료의 수는 Table 2.3과 같다. 3세인 경주마와 4세인 경주마가 각각 전체 자료의 40.51%, 30.62%로 가장 많은 수를 차지한다.

  다음으로 경주마의 연령과 순위 정보를 이용하여 각 연령 별로 순위가 1–3위에 해당하는 비율을 계산해본 결과 Figure 2.1과 같다. 그래프를 통해 대체적으로 경주마의 연령이 낮을수록 3위안에 들어오는 비율이 높다는 것을 알 수 있다.

- 부담중량

  부담중량이란 경주에서 말이 등에 짊어지고 달리는 무게를 말하는 것으로, 부담중량에 포함되는 것은 기수체중, 웨이트패드 등 중량조정 물품, 기수장구, 안장 등이며 부담중량에 포함되지 않는 것은 기수채찍, 안전모, 보호조끼, 말 등번호판 등이다. 부담중량은 경주마의 연령, 성별, 산지, 경주성적 등에 따라 다양한 방법으로 부여할 수 있고, 크게 마령중량, 별정중량, 핸디캡중량 3가지로 분류된다. 일반적으로 부담중량 1kg의 차이는 시간으로 1/3초, 거리로는 약 4.8m(2마신)이기 때문에 부담중량은 순위를 예측하는데 중요한 변수라고 할 수 있다.

- 마체중, 증감

  마체중은 경주마의 체중(kg)으로 수집한 자료에서는 370kg∼585kg까지 존재하며 평균체중은 469.9 kg이다.

  증감은 전 경주와 비교한 경주마의 체중 변화(kg)로 −46kg부터 +41kg까지 존재하며 보통 1일 체동변동범위는 10∼15kg이다. 체중의 감소요인으로는 많은 운동량과 컨디션 불량으로 인한 급식상태 불

량이 있으며, 증가요인으로는 운동량의 부족, 성장기의 자연성장, 계절요인인 가을 등이 있다. 2~3세의 성장기의 말을 제외하고 경주마의 체중은 지난 출주 대비 보통 5kg 정도 늘거나 줄며 경주 성적에 특히 영향을 미치는 경우는 10~12kg 이상의 변화를 보일 때이다. 일반적으로 0~5kg 선에서 체중이 변화한 경주마의 승률이 가장 높고, 체중의 변화가 10kg 이상이면 우승마 선정 시 주의해야 한다.

- 등급

  경주마는 능력에 따라 적정한 군(群)에 편성된다. 일반적으로 최하위 군부터 시작하여 경주성적에 따라 상위 군으로 승군하게 되며 상위 군일수록 상금이 많아진다. 원 자료에서는 국내산마와 외국산마, 1군~6군으로 구분되어 있다. 두 가지 정보를 모두 사용하기 위해 국외(국내/해외)변수와 군(1~6)변수로 나누었고, 군 변수는 그 수가 작을수록 좋은 등급의 말을 의미한다.

- 말 1위 비율, 말 2위 비율

  원 자료에서는 경주마의 총 출전 횟수와 총 1위 횟수, 총 2위 횟수 정보가 존재한다. 하지만 총 출전 횟수가 많을수록 총 1위와 2위 횟수도 많은 경향이 있기 때문에 이를 보정하기 위해 말 1위 비율과 말 2위 비율로 데이터를 변환하였다. 따라서 말 1위 비율은 경주마의 총 출전 횟수 중 총 1위 횟수의 비율이고, 말 2위 비율은 경주마의 총 출전 횟수 중 총 2위 횟수의 비율이다. 다음에 나오는 기수정보와 조교사정보에서도 마찬가지로 횟수 정보를 비율 정보로 변환하였다.

- 말 1년 출전 횟수, 말 1년 1위, 말 1년 2위

  말 1년 출전 횟수는 경주마의 최근 1년간 출전 횟수, 말 1년 1위는 경주마의 최근 1년간 1위 횟수, 말 1년 2위는 경주마의 최근 1년간 2위 횟수를 의미한다. 경주마의 최근 정보는 순위를 예측하는데 중요한 정보로 여겨지기 때문에 예측모형의 주요변수라고 할 수 있다.

2.3.4. 조교사 관련 정보 말의 선천적 능력 이외에 조교사의 훈련을 통한 후천적 능력 또한 경주마의 순위에 중요한 영향을 미칠 수 있다. 따라서 조교사의 나이와 경력, 훈련시킨 경주마의 1위 비율, 2위 비율, 최근 1년간 총 출전 횟수, 최근 1년간 1위 횟수, 최근 1년간 2위 횟수 정보가 포함되었다.

2.3.5. 기수 관련 정보 기수의 능력 또한 경주마의 순위 예측에 영향을 미칠 수 있다. 따라서 기수의 나이와 경력, 1위 비율, 2위 비율, 최근 1년간 총 출전 횟수, 최근 1년간 1위 횟수, 최근 1년간 2위 횟수 정보가 포함되었다. 생성된 변수들을 정리한 결과는 Table 2.4와 같다.

3. 분석 결과

이번 장에서는 서울시 경마 경기의 우승마를 예측하기 위하여 1) 순위를 기반으로 한 모형, 2) 기록을 기반으로 한 모형을 적합하여 경기별 우승마 예측률을 비교해보고, 예측 모형에 포함된 중요한 변수에 대하여 알아본다.

2014년 1월부터 2015년 4월까지의 자료인 총 1474회의 경기 중 임의로 70%의 경기를 train data로 30%의 경기를 test data로 나누어 test data에서의 예측률을 모형 비교의 지표로 사용하였다. 위와 같은 과정을 100회 반복 시행하여 평균 예측률을 비교하고 최적 모형을 구하고자 한다. 그리고 2015년 5월 한 달간 자료인 109회의 경기에 최적 모형을 적용하여 도출한 예측률과 실제 배당률을 적용한 배당 금액을 계산하여 비교한다. 우리는 분석에서 중요변수 도출과 최종모형 선택은 다음과 같이 수행하였다. 우선 선형모형의 경우에는 일단 마지막 1달치 데이터를 제외한 모든 데이터를 사용해서 선형모형을 적합하고 AIC/BIC 값을 최소화 하는 최적의 모형을 구하고 이 때 선택된 설명변수들을 저장한다.

Table 2.4. Description of variables

	Variable	Description	Type
Input variables	주로 습도	경주가 시행된 시점의 주로 습도(%)	Numerical
	마번	경주시 경주마에게 부여되는 번호	
	연령	경주마의 연령	
	부담중량	경주마가 부담해야 하는 중량으로 기수의 체중, 장구, 안장, 안장모포 등의 무게	
	마체중	경주마의 체중(kg)	
	증감	전 경주와 비교한 경주마의 체중 변화(kg)	
	말1위비율	경주마의 총 1위 횟수/경주마의 총 출전 횟수	
	말2위비율	경주마의 총 2위 횟수/경주마의 총 출전 횟수	
	말1년출전횟수	경주마의 최근 1년간 출전 횟수	
	말1년1위	경주마의 최근 1년간 1위 횟수	
	말1년2위	경주마의 최근 1년간 2위 횟수	
	군	경주마가 속해있는 군으로 1군 6군으로 구분	
	조교나이	조교사의 나이	
	조교경력	조교사의 경력(년)	
	조교1위비율	훈련시킨 경주마의 총 1위 횟수/훈련시킨 경주마의 총 출전 횟수	
	조교2위비율	훈련시킨 경주마의 총 2위 횟수/훈련시킨 경주마의 총 출전 횟수	
	조교1년출전횟수	훈련시킨 경주마의 최근 1년간 총 출전 횟수	
	조교1년1위	훈련시킨 경주마의 최근 1년간 총 1위 횟수	
	조교1년2위	훈련시킨 경주마의 최근 1년간 총 2위 횟수	
	기수나이	기수의 나이	
	기수경력	기수의 경력(년)	
	기수부담중량	기수의 체중(kg)	
	기수1위비율	기수의 총 1위 횟수/기수의 총 출전 횟수	
	기수2위비율	기수의 총 2위 횟수/기수의 총 출전 횟수	
	기수1년출전횟수	기수의 최근 1년간 출전 횟수	
	기수1년1위	기수의 최근 1년간 1위 횟수	
	기수1년2위	기수의 최근 1년간 2위 횟수	
	성별	경주마의 성별로 암컷, 수컷, 거세로 구분	Categorical
	국외	경주마가 속해있는 군으로 국내/해외로 구분	
Response variables	순위	경주 별 경주마가 들어온 순서	Categorical
	기록	경마 경주시의 기록(0.0초)	Numerical

그리고 train/test로 나누어서 train data에서 이 설명변수들을 사용해서 모형을 적합한다. 물론 train data는 매번 random하게 선택되므로 회귀계수 값들은 달라지지만 사용된 설명변수는 동일하다. 마지막 1달치 데이터에 적용한 최종 선형 모형은 100개의 test data에서 test error가 가장 적은 모형을 선택한다. 랜덤 포레스트의 경우는 다른 방법을 이용하였다. 랜덤포레스트에서는 OOB error를 제공하므로 100개의 train set에서 적합된 모형 중에서 OOB error가 가장 적은 모형을 최적의 모형으로 선택한다. 그리고 이 최적 모형에서 산출된 variable importance를 기반으로 중요변수를 도출한다. 마지막 1달치 데이터에 적용된 최종모형은 100개의 test data에서 test error가 가장 적은 모형을 선택한다.

3.1. 순위를 기반으로 한 모형

순위를 기반으로 한 예측 모형은 경마의 특성상 배팅 방식에 따라 예측해야하는 말의 수가 다르므로 1위인 한 마리를 예측하는 경우, 2위까지 두 마리, 3위까지 세 마리를 예측하는 모형으로 각각 다르게 적합

Table 3.1. The important variables of each logistic model

	선택된 변수 (+)	선택된 변수 (−)
1마리	군, 기수1년1위, 기수1위비율, 마체중, 말1년1위, 말1위비율, 말2위비율	기수1년출전횟수, 마번, 말1년출전횟수, 성별수, 성별암, 증감
2마리	국외외, 군, 기수1년1위, 기수2위비율, 마체중, 말1년1위, 말1년2위, 말1위비율, 말2위비율	기수1년출전횟수, 마번, 말1년출전횟수, 부담중량, 성별암
3마리	군, 기수1년1위, 기수2위비율, 말1년1위, 말1년2위, 말1위비율	기수1년출전횟수, 마번, 말1년출전횟수, 성별수, 성별암, 연령

Table 3.2. The important variables of each random forest model

	주요변수 10개
1마리	기수1년1위, 기수1위비율, 기수2위비율, 마체중, 말1위비율, 조교1년출전횟수, 조교1위비율, 조교2위비율, 증감, 기수1년출전횟수
2마리	기수1년출전횟수, 기수1위비율, 기수2위비율, 마체중, 말1년출전횟수, 말1위비율, 말2위비율, 조교1년출전횟수, 조교1위비율, 조교2위비율
3마리	기수1년출전횟수, 기수1년1위, 기수1위비율, 기수2위비율, 마체중, 말1위비율, 말2위비율, 조교1년출전횟수, 조교1위비율, 조교2위비율

하였다. 예를 들어 n위까지 n마리를 예측하는 모형에서는 순위 값이 1부터 n위인 말의 순위를 1로 바꾸고, 이하 순위의 말은 모두 순위를 0으로 바꾸어 2-class 분류 문제로 변환하여 분석하였다. 순위는 한 경기 내에서 결정되는 값이므로 같은 경주를 뛰는 경주마 간에 차이가 없는 주로 습도 변수는 제외하고 분석하였다.

분류분석방법으로 모든 변수를 전부 이용한 로지스틱 회귀와 AIC, BIC 기준으로 변수선택법 (Park 등, 2011; Venables와 Ripley, 2002)을 이용한 로지스틱 회귀, 그리고 랜덤 포레스트 방법을 이용하여 test data에서의 예측률을 비교하였다.

위의 방법으로 순위를 0, 1로 바꾸면 1보다 0이 훨씬 많은 불균형 데이터이므로 적합시킨 분류 모형으로 test data를 분류하면 각 경기마다 원하는 마릿수만큼 예측이 되지 않는 경우가 있다. 예를 들어, 어느 경기에는 1등 말이 하나도 없는 경우가 나올 수도 있다. 따라서 경기 별로 배팅 방식에 따른 마릿수를 예측하기 위하여 개별 말이 1로 분류 될 예측 확률을 계산하였다. 이후 n마리의 우승마를 예측하기 위해서 각 경기 내에서 1로 분류될 예측 확률이 높은 순서대로 n위까지의 말을 우승마로 하여 1로 나머지를 0으로 분류하였다.

먼저 로지스틱 회귀 모형에 대해서 살펴보면, BIC 기준 로지스틱 모형에서 선택된 변수는 모두 AIC 기준에서 선택된 변수에 포함된다. 그러므로 AIC 기준의 로지스틱 모형에서 선택된 변수 중 유의 수준 0.01하에서 유의한 변수의 부호만을 살펴보았으며 이는 Table 3.1과 같다. 로지스틱 회귀에서 회귀계수가 양수이면 설명변수의 값이 커질수록 우승마가 될 확률이 증가하는 변수임을 의미한다. 양의 효과를 갖는 변수를 살펴보면, 마체중이 무거울수록, 기수와 말의 우승 비율이 높을수록 우승마일 확률이 높아지는 것을 확인할 수 있다. 음의 효과를 갖는 변수를 살펴보면, 마번이 커질수록 우승확률이 낮아지므로 마번이 작은, 즉 안쪽에서 출발하는 것이 유리하다는 사실을 확인할 수 있다. 말의 성별이 수나 암이면 우승할 확률이 작아지며, 회귀계수를 통해 거세마, 수말, 암말 순으로 빠르다는 것을 알 수 있다.

다음은 랜덤 포레스트에 대해서 살펴보고자 한다. 랜덤 포레스트 모형의 경우 로지스틱 회귀에서와 달리 회귀계수의 값이 주어지지 않으므로 변수 중요도가 높은 상위 10개 변수를 살펴보았으며, 이는 Table 3.2에서 확인할 수 있다. 로지스틱 회귀 모형에서와 유사하게 말의 과거 우승 비율, 기수의 우승 비

Table 3.3. Average prediction accuracy in test data

	Logistic	Logistic-AIC	Logistic-BIC	Random Forest
1마리	0.4269 (0.0197)	0.4322 (0.0187)	0.4303 (0.0179)	0.3777 (0.0214)
2마리	0.2357 (0.0167)	0.2367 (0.0177)	0.2317 (0.0159)	0.2153 (0.0184)
3마리	0.0894 (0.0123)	0.0898 (0.0124)	0.0923 (0.0114)	0.0945 (0.0131)

Table 3.4. The important variables of each linear model

거리	R^2 (Adj-R^2)	선택된 변수 (+)	선택된 변수 (−)
1000m	0.3727 (0.3660)	군, 기수나이, 마번, 부담중량, 성별수, 성별암, 연령, 조교나이	기수경력, 기수1년1위, 말1년출전횟수, 말1위비율, 말2위비율, 조교경력, 조교1년1위
1100m	0.2108 (0.1877)	군, 기수부담중량, 연령	말1위비율, 말2위비율
1200m	0.4308 (0.4284)	군, 마번, 부담중량, 성별수, 성별암, 조교나이	국외외, 기수2위비율, 말1년출전횟수, 말1위비율, 말2위비율, 조교경력, 조교1년출전횟수
1300m	0.4309 (0.4283)	군, 마번, 부담중량, 성별암, 조교나이	기수1위비율, 기수1년2위, 말1년출전횟수, 말1위비율, 말2위비율, 조교경력, 조교1년1위, 주로 습도
1400m	0.4120 (0.4089)	군, 마번, 부담중량, 성별암, 증감	국외외, 기수1년1위, 말1년출전횟수, 말1위비율, 말2위비율, 조교1년출전횟수, 조교2위비율, 주로 습도
1700m	0.3046 (0.2900)	군, 성별암, 주로 습도	기수1년2위, 조교1년출전횟수, 말1년출전횟수, 말1위비율, 말2위비율
1800m	0.4979 (0.4942)	군, 기수경력, 기수1년출전횟수, 성별암, 주로 습도, 증감	국외외, 기수1년1위, 기수2위비율, 말1년1위, 말1위비율
1900m	0.3537 (0.3356)	군	국외외, 기수1년1위, 마체중, 말1년1위, 말1년2위, 성별암, 주로 습도
2000m	0.5010 (0.4725)	군, 성별암, 연령, 조교경력	기수2위비율, 말1년출전횟수, 말1위비율, 부담중량, 조교1년출전횟수, 주로 습도

율, 마체중 등이 중요한 변수로 선택된 것을 알 수 있다.

각 모형에 대한 test data에서의 평균 예측률(표준편차)을 계산한 결과는 Table 3.3에서와 같다. 1마리, 2마리 예측에서는 AIC-로지스틱 모형이, 3마리 예측 모형에서는 랜덤 포레스트 모형의 예측률이 가장 높았다. 하지만 1, 2마리 예측에서 AIC-로지스틱 모형과 BIC-로지스틱 모형의 예측력이 거의 같고, 3마리 예측에서는 BIC-로지스틱 모형이 더 우수하였다. 즉, BIC-로지스틱 모형이 AIC-로지스틱 모형보다 더 간단하고 예측력은 거의 차이가 없으므로 BIC-로지스틱 모형을 최적 모형으로 선택하였다. 그러나 실제 배당률을 적용하여 시행하는 경우에는 어떤 모형이 더 좋은 결과를 나타낼지 알 수 없으므로 3.3절에서 AIC-로지스틱 모형과 BIC-로지스틱 모형을 모두 비교해보고자 한다. 랜덤 포레스트의 최적 모형은 100번 적합 시 train data에서의 예측률을 최대로 하는 모형으로 하였다.

3.2. 기록을 기반으로 한 모형

기록을 기반으로 한 모형은 원하는 마릿수의 우승마를 예측하기 위하여 거리에 따라 각각 회귀 모형을 적합한 후, 경기 별로 기록에 따른 순위를 부여하였다. 기록 모형에서도 순위에 기반한 예측 모형과 같이 배팅 방식에 따라 예측률을 계산하여 비교하였다. 회귀분석방법으로는 AIC 기준의 단계별 변수선택

Table 3.5. The important variables of each random forest model

거리	R^2 (Adj-R^2)	주요변수 10개
1000m	0.3644 (0.3544)	국외, 군, 기수1위비율, 기수2위비율 , 마체중, 말1년출전횟수, 말1위비율, 말2위비율, 조교1위비율, 증감
1100m	0.1563 (0.0633)	군, 기수1년출전횟수, 마체중, 말1년출전횟수, 말1위비율, 말2위비율, 조교1위비율, 조교2위비율, 주로 습도, 증감
1200m	0.4644 (0.4610)	군, 기수1년1위, 기수1위비율, 기수2위비율, 마체중, 말1년1위, 말1위비율, 말2위비율, 조교2위비율, 증감
1300m	0.4187 (0.4140)	군, 말1위비율, 말2위비율, 최근1년1위, 마체중, 증감, 기수2위비율, 주로 습도, 조교1위비율, 조교2위비율
1400m	0.4409 (0.4361)	군, 기수1위비율, 기수2위비율, 마체중, 말1년1위, 말1위비율, 말2위비율, 조교1위비율, 주로 습도, 증감
1700m	0.5190 (0.4997)	군, 기수1년1위, 기수1년출전횟수, 기수1위비율, 기수2위비율, 말1위비율, 말2위비율, 조교1년출전횟수, 주로 습도, 증감
1800m	0.6570 (0.6517)	국외, 군, 기수1년1위, 기수1위비율, 기수2위비율, 마체중, 말1위비율, 말2위비율, 주로.습, 증감
1900m	0.5678 (0.5374)	군, 기수1년1위, 기수1년출전횟수, 기수1년2위, 기수1위비율, 기수2위비율, 말1위비율, 부담중량, 조교1년2위, 주로 습도
2000m	0.5715 (0.5250)	기수1년1위, 기수1위비율, 기수2위비율, 말1위비율, 말1년1위, 부담중량, 조교1년출전횟수, 주로 습도, 조교1년2위, 조교2위비율

법을 이용한 선형 회귀와 랜덤 포레스트 방법을 이용하였다. BIC 기준의 단계별 변수선택법도 시행하였으나 AIC 기준의 단계별 변수선택법을 이용한 결과와 거의 동일하여 AIC 기준의 모형을 이용하기로 한다.

단계별 변수선택법을 사용하여 적합한 경주 거리별 예측 모형에서 유의수준 0.05하에서 선택된 변수는 Table 3.4와 같다. 결정계수 R^2의 최솟값은 0.2108, 최댓값은 0.4979으로 나타났다. 선형 회귀에서 회귀계수의 부호가 음수이면 경주마의 기록 단축에 영향을 미치는 변수임을 의미한다. 먼저 회귀계수가 양수인 설명변수를 살펴보고자 한다. 군 변수는 모든 경주 거리별 모형에 포함되므로 중요한 변수이며, 그 값이 작을수록 좋은 경주마임을 확인할 수 있다. 성별암 변수에 의하면 1900m를 제외하고는 암말이 수말이나 거세마에 비해 기록이 느리다는 것을 알 수 있다. 부담중량은 대부분의 경우 회귀계수가 양수이나 2000m에서는 음수이다. 이는 연령, 성별, 최근 승군점수 등에 따라 능력이 좋은 말에게 높은 부담중량을 부여하는 경우도 있기 때문이다. 회귀계수가 음수인 설명변수는 말 1위 비율이 1900m 모형을 제외한 모든 예측 모형에 포함되었고, 말 2위 비율은 6개의 모형에, 말 1년 출전 횟수는 5개의 모형에 포함되어 주요변수라고 할 수 있다. 따라서 말의 1, 2위 비율이 높을수록, 최근 1년 출전횟수가 많을수록 기록이 좋아지는 것을 알 수 있다.

다음은 랜덤 포레스트에 대해서 살펴보고자 한다. 앞 절에서와 같이 랜덤 포레스트 모형에서의 각 거리별로 상위 10개의 주요변수를 살펴보았으며, 이는 Table 3.5와 같다. 결정계수 R^2의 최솟값은 0.1563, 최댓값은 0.6570으로 나타났다. 설명변수 중 말 1위 비율은 거리별 주요변수에 모두 포함되었으며, 군과 기수 2위 비율은 하나의 거리모형을 제외하고는 모두 포함되었다. 그 외에 거리별 모형의 주요변수는 말 2위 비율, 기수 1위 비율, 증감, 주로 습도 변수 순으로 선택되었다. 이를 통해 말과 기수의 정보와 관련된 설명변수가 기록을 예측함에 있어 중요한 것을 알 수 있었다. 또한 경주로의 습도가 주요변수이므로 말이 경주하는 환경이 기록에 영향을 미치는 것을 알 수 있다.

Table 3.6. Average prediction accuracy in test data

예측률	Linear model			Random Forest model		
거리	1마리	2마리	3마리	1마리	2마리	3마리
1000m	0.4728 (0.0622)	0.3532 (0.0540)	0.1160 (0.0380)	0.5084 (0.0691)	0.3384 (0.0587)	0.1398 (0.0436)
1100m	0.5200 (0.1610)	0.2322 (0.1129)	0.1300 (0.0961)	0.2911 (0.1261)	0.1044 (0.1045)	0.0444 (0.0688)
1200m	0.4772 (0.0402)	0.3364 (0.0365)	0.1285 (0.0224)	0.5172 (0.0419)	0.2878 (0.0314)	0.1310 (0.0316)
1300m	0.4650 (0.0416)	0.3078 (0.0385)	0.1537 (0.0326)	0.4503 (0.0488)	0.2812 (0.0419)	0.1339 (0.0340)
1400m	0.4577 (0.0463)	0.2765 (0.0424)	0.0899 (0.0272)	0.4370 (0.0519)	0.2576 (0.0503)	0.1002 (0.0291)
1700m	0.4019 (0.0846)	0.1119 (0.06290)	0.1062 (0.06010)	0.3495 (0.0974)	0.1162 (0.0618)	0.0657 (0.0514)
1800m	0.4129 (0.0658)	0.2210 (0.0494)	0.1058 (0.0371)	0.3848 (0.0568)	0.2165 (0.0522)	0.1010 (0.0399)
1900m	0.2958 (0.1055)	0.2250 (0.1127)	0.1042 (0.0815)	0.2667 (0.1236)	0.1675 (0.1022)	0.1167 (0.0991)
2000m	0.3250 (0.1293)	0.0700 (0.0911)	0.1425 (0.1138)	0.4300 (0.1592)	0.1225 (0.1164)	0.2000 (0.1281)
평균	**0.4537**	**0.2888**	**0.1214**	**0.4528**	**0.2614**	**0.1197**

단계별선택법을 이용하여 적합한 예측 모형과 랜덤 포레스트를 이용하여 적합한 예측 모형에서 공통적으로 선택된 주요변수는 말 1위 비율과 말 2위 비율, 군이다. 따라서 기록을 기반으로 경주마의 순위를 예측하기 위해서는 말과 관련된 정보가 중요하게 작용한다는 사실을 확인할 수 있다.

각 거리 별 예측 모형에서 test data에서의 평균 예측률(표준편차)을 계산한 결과는 Table 3.6에서와 같다. 대부분 예측률은 1마리를 예측하는 경우가 가장 높고 2마리, 3마리 순으로 높다. 하지만 예외적으로 2000m의 경우에는 2마리를 예측하는 경우보다 3마리를 예측하는 경우의 예측률이 더 높게 나왔다. 이는 상위 3등의 기록 예측이 3-2-1순으로 예측되거나 2-3-1순으로 예측되어 1, 2위 2마리 예측에는 실패하였지만, 1, 2, 3위 3마리 예측은 정확히 하여 나타난 결과이다. 거리별 예측률을 각 경기 수에 따라 가중치 평균을 낸 결과, 모든 경우 가중치 평균 예측률이 랜덤 포레스트 모형보다 선형 회귀 모형에서 높게 나온 것을 알 수 있다. 랜덤 포레스트 모형의 최적 모형은 순위를 기반으로 한 모형에서와 같은 방식으로 하였다.

3.3. 배당률 이용 결과

앞 절에서 선택된 최적모형을 이용하여 2015년 5월 한 달 동안의 test data(총 109개의 경기)에 적용하였을 때의 예측률을 비교하였다. 1마리 예측 모형은 단승식의 경우에 해당하고, 2마리, 3마리 예측 모형은 각각 복승식, 삼복승식에 해당하므로 배팅 방식에 따른 배당률을 이용하여 얻을 수 있는 배당 금액을 계산하였다.

Table 3.7을 통해 전체적인 결과를 살펴보면 다음과 같은 사실을 발견할 수 있다. 단승식의 경우에는 기록을 기반으로 한 선형회귀모형의 예측률이 가장 높고, 복승식의 경우에는 기록을 기반으로 한 랜덤포레스트 모형의 예측률이 가장 높았다. 그러나 삼복승식의 경우에는 순위를 기반으로 한 BIC-로지스틱 모

Table 3.7. Prediction accuracy of each model

	예측률			
	순위를 기반으로 한 모형		기록을 기반으로 한 모형	
	logistic-BIC	Random Forest	Linear	Random Forest
단승식	0.4220	0.3853	0.4312	0.4037
복승식	0.2110	0.2110	0.2110	0.2385
삼복승식	0.1651	0.1284	0.1468	0.1193

Table 3.8. Profits of each model (단위: 원)

	이윤 금액			
	순위를 기반으로 한 모형		기록을 기반으로 한 모형	
	logistic-BIC	Random Forest	Linear	Random Forest
단승식	2,328,000	2,344,000	2,830,000	2,643,000
복승식	612,000	518,000	429,000	7,996,000
삼복승식	24,813,000	550,000	23,465,000	22,431,000
총합	**27,753,000**	**3,412,000**	**26,724,000**	**33,070,000**

Table 3.9. Profits of Model by distance (단위: 원)

	거리별 이윤 금액			
	순위		기록	
	logistic-BIC	Random Forest	Linear	Random Forest
1000m	24,448,000	2,247,000	24,451,000	31,822,000
1100m	-	-	-	-
1200m	2,044,000	228,000	613,000	700,000
1300m	274,000	329,000	16,000	530,000
1400m	916,000	696,000	1,853,000	319,000
1700m	299,000	−68,000	7,000	−45,000
1800m	−189,000	19,000	−177,000	−217,000
1900m	-	-	-	-
2000m	−39,000	−39,000	−39,000	−39,000
총합	27,753,000	3,412,000	26,724,000	33,070,000

형의 예측률이 가장 높은 결과를 보였다.

다음으로 배당률 정보를 이용하여 각 경기당 10,000원씩 배팅하였을 때의 총 이윤금액을 계산해보았다. 우승마를 맞췄을 경우에는 10,000원 $*$ (배당률 $-$ 1), 맞추지 못했을 경우에는 $-$10,000원으로 한 것을 합하여 한 달간 총 이윤 금액을 계산하였다. 세금을 공제하기 전의 이윤 금액은 Table 3.8과 같다. 이윤 금액은 예측률 결과와 같이 단승식에서는 기록을 기반으로 한 선형회귀모형이, 복승식에서는 기록 기반 랜덤 포레스트 모형이, 삼복승식에서는 BIC-로지스틱 모형이 가장 높았다. 이윤 금액의 합이 가장 큰 모형은 기록을 기반으로 한 랜덤 포레스트 모형인 것을 확인할 수 있다.

Table 3.9를 통해 거리별 이윤 금액을 자세히 살펴보면 1000m에서의 이윤 금액이 다른 거리에 비해 월등히 크다는 것을 알 수 있다. 이는 33번째 경기(2015.5.10, 1000m)의 배당률이 각각 233.2(단승식), 731(복승식), 2221(삼복승식)으로 매우 높기 때문에 이 경기를 맞춤으로써 높은 이윤 금액을 얻게 된 것이다. 따라서 실제 이윤 금액에 큰 영향을 미치는 것은 이러한 배당률이 높은 경기를 맞출 수 있는지 여부이다.

4. 결론

본 연구에서는 경마 경기의 우승마 예측을 위해 한국 마사회에서 제공하는 경마 성적표, 경주마 정보, 기수와 조교사 정보를 사용하여 예측 모형을 제시하였다. 예측 모형은 순위를 기반으로 한 예측 모형과 기록을 기반으로 한 예측 모형을 적합하였고 각 예측 모형의 주요 변수를 살펴보았다. 최적 예측 모형의 정확도를 비교하기 위하여 단승식과 복승식, 삼복승식에 따른 예측률을 평가 지표로 사용하였다.

순위를 기반으로 한 예측 모형은 분류분석방법을 사용하여 모든 변수를 사용한 로지스틱 회귀모형과 AIC, BIC를 기준으로 단계별 변수선택법을 이용한 로지스틱 회귀모형, 랜덤 포레스트 모형의 총 4가지 예측 모형을 적합하였다. 그 결과, 단계별 변수선택법을 이용한 회귀 모형을 통해 마체중이 높을수록, 기수와 말의 우승 비율이 높을수록 우승마일 확률이 높아지는 것을 확인할 수 있었다. 또한, 거세마가 암말이나 수말보다 빠르며, 예상한 바와 같이 마번이 작은 안쪽에서 출발하는 것이 유리하다는 사실을 확인할 수 있었다. 랜덤 포레스트 예측 모형의 경우, 주어진 설명변수의 중요도를 통해 말과 기수의 과거 우승 비율이 순위 예측에 주요 역할을 하는 것을 알 수 있었다. 순위를 기반으로 한 각 예측 모형의 예측률을 비교해본 결과, BIC-로지스틱 모형과 랜덤 포레스트 모형을 최적 모형으로 선택하였다.

기록을 기반으로 한 예측 모형은 각 경주 거리별로 모형을 적합하였으며, 적합 시 단계별선택법과 랜덤 포레스트를 이용하였다. 단계별선택법을 이용한 선형회귀 모형에서는 군, 성별암, 부담중량 등의 변수가 기록 증가에 영향을 미치는 것을 알 수 있었다. 반면, 기록 단축에 영향을 미치는 유의한 설명변수로는 말 1, 2위 비율, 말 1년 출전 횟수 등의 변수가 선택되었다. 랜덤 포레스트를 이용한 예측 모형의 경우도 선형회귀모형과 비슷하게 말 1, 2위 비율, 군 등 말에 관한 정보가 주요변수로 선택되었다. 또한 기수 1, 2위 비율, 기수 1년 1위 등 기수의 과거 우승 경력이 기록에 영향을 미치는 것으로 나타났다. 그리고 주로 습도와 같이 경주 환경과 직접적으로 관련이 있는 설명변수도 경마 기록에 영향을 미치는 것을 알 수 있었다. 두 모형에 공통적으로 선택된 주요변수는 말 1, 2위 비율, 군 변수로 말에 관한 정보가 기록 예측에 중요한 역할을 하는 것을 알 수 있었다. 두 모형의 예측률를 비교한 결과, 근소한 차이로 선형회귀 모형이 랜덤 포레스트 모형보다 단승식과 삼복승식에서 더 나은 예측률을 보였다.

순위를 기반으로 한 예측 모형과 기록을 기반으로 한 예측 모형의 예측률을 비교해 보았을 때, 비슷한 예측률을 보인다는 것을 알 수 있었다. 또한 단승식과 복승식, 삼복승식의 배당률 정보를 이용하여 획득하게 되는 이윤금액을 살펴본 결과, 이윤 금액은 특정 경기의 배당률에 큰 영향을 받는 사실을 확인할 수 있었다.

본 연구에서 제시한 예측 모형들은 데이터마이닝 기법을 이용하여 통계적 분석에 근거한 모형으로, 우승마 예측에 있어 임의로 선택하는 경우보다 훨씬 높은 예측률을 보인다. 따라서 앞으로의 경마 경기에 대한 정보가 충분히 제공된다면 이러한 예측 모형들이 우승마 예측에 도움을 줄 수 있으리라 기대되는 바이다.

References

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Hastie, T. J. and Pregibon, D. (1992). *Generalized Linear Models*, Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, **37**, CRC press.

Park, C., Kim, Y., Kim, J., Song, J. and Choi, H. (2011). *Datamining using R*, Kyowoo, Seoul.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org

Statistics Korea e-National indicators (2015). http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx

_cd=1662

The Korea Racing Authority (2014). http://www.kra.co.kr/main.do

The National Gambling Control Commission (2015). http://static.ngcc.go.kr/user/index.jsp

The National Gambling Control Commission (2014). http://www.ngcc.go.kr/Board/ReadView.do?idx=pds&page=1&no=9346

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, Springer, New York.

Yoo, S. and Park, H. (2000). The horse race winning probability via logistic regression, *Korean Journal of Applied Statistics*, **13**, 35–44.

서울 경마 경기 우승마 예측 모형 연구

최혜민[a] · 황나영[a] · 황찬경[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2015년 9월 14일 접수, 2015년 10월 13일 수정, 2015년 10월 19일 채택)

요 약

경마 산업은 국내 합법 사행산업의 대부분을 차지하고 있다. 그러나 사행성 도박이라는 인식 하에 여타 스포츠 산업에 비해 활발한 통계적 분석이 이루어지지 않고 있다. 본 연구의 목적은 다양한 데이터마이닝 기법을 이용하여 우승마를 예측하는 모형 개발에 있다. 모형 적합에 사용한 데이터는 한국 마사회에서 제공하는 자료를 바탕으로 하였으며, 경마 성적표, 경주마 정보, 기수 정보, 조교사 정보 등을 사용하였다. 예측 모형은 크게 두 모형으로 나누어 순위를 기반으로 한 모형과 기록을 기반으로 한 모형으로 적합하였고, 분석 방법으로는 선형회귀분석, 랜덤 포레스트, 로지스틱 회귀 분석을 사용하였다. 그 결과 말 기본 정보와 과거 우승 경력, 기수의 과거 우승 경력 등이 순위 예측에 큰 영향을 미치는 것을 알 수 있었다. 모형 적합에 사용되지 않은 최근 1개월 간 데이터를 이용하여 단승식, 복승식, 삼복승식으로 배팅한 결과 모형 간 큰 차이가 없었고, 모두 양의 수익을 얻을 수 있었다.

주요용어: 경마, 선형회귀모형, 단계적 회귀분석, 랜덤 포레스트, 로지스틱 회귀분석, 주요변수

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2013R1A1A2012817).

[1]교신저자: (03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr