응용통계연구 제 **13**권 *1*호, *2000*년, *pp. 35-43*

로지스틱 회귀를 통한 경마의 입상확률모형

유선경[1] 박홍선[2]

요 약

본 연구에서는 우리 나라 경마의 실제자료를 이용하여 연승식 경마의 입상확률에 미치는 여러 가지 요인을 조사하였고, 이를 토대로 입상확률모형을 유도하여 보았다. 외국의 경우, 경마에 대한 통계적 접근이 다각적으로 시행되었지만, 기존의 선행방법이 배당금에 의한 입상확률에 근거를 하고 있는 반면, 본 연구에서는 경마장에서 쉽게 구할 수 있는 정보를 중심으로, 로지스틱 회귀를 이용한 방법을 시도해 보았다.

주요용어: 동질성 검정, 단계별 선택, 분할표.

1. 서론

경마(Horse Race)는 우리 나라에 1898년경에 처음으로 소개되었고, 1920년경에 본격적인 운영이 이루어진 것으로 알려져 있다. 그 후, 경마는 경제발전과 더불어 1970년대에 비약적 발전을 하였으나, 단일마주제(경주마 소유가 마사회)로 인한 승부조작으로 그 투명성과 객관성이 의심되어 왔다. 최근 1993년 개인마주제(누구든지 경주마를 소유할 수 있는 제도)의 발족으로 경마자료의 투명성과 객관성은 정착되기 시작했다. 그럼에도 불구하고, 우리 나라에서는 경마자료에 대한 통계적 분석방법에 대한 선행연구가 없었기에 본 연구는 경마자료의 통계적 접근을 시도하였다.

경마는 영국, 호주, 홍콩, 그리고 미국, 일본 등 선진국에 널리 정착된 고급 여가 문화로 여겨지고 있고, 그 만큼, 입상확률에 대한 많은 연구가 있어 온 것도 사실이다. (여기서 입상확률-winning probability-이란 자신이 배팅한 경마가 입상할 확률을 의미한다.) 경마 자료가 사람들의 흥미를 불러일으키는 이유는 다음과 같다. 첫째, 많은 사람이 쉽게 참여할 수 있다는 점이고, 둘째, 경주에 관련된 수많은 정보가 존재한다는 점이다. 그 뿐 아니라, 다른 경주(경차, 경륜 등)와 마찬가지로 여러 선수가 한꺼번에 승부(multi-entry competition)를 겨루는 복잡한 확률 문제이기도 하면서, 이긴 말에 건 소득액을 배팅한 사람에게 수수료를 제외하고 돌려주는 pari-mutuel 방식으로 진행되기 때문에, 경제학자들은 경마의 이론을 내기시장(wagering market)의 효율성(efficiency)을 평가하는데 활용하고 있기도 하다(Ali, 1979, 1994). 또한, Snyder (1978)는 불확실성과 위험(risk)을 고려한 상태의 결정(decision)에 대해 그 평가의 기회가 가능하다는 점에서 경마를 증권이나 채권 같은 유가증권시장에 비유하고 있다.

1) (151-742) 서울 관악구 신림동 산 56-1, 서울대학교 통계학과 석사과정
2) (449-791) 경기도 용인시 모현면 한국외국어대학교 통계학과 부교수
E-mail: hspark@stat.hufs.ac.kr

지금까지 입상확률에 대한 모형에서 가장 널리 사용되는 것은 Harville(1973) 의 모형으로, 경주에 임하는 i-번째 경주마가 1등이 될 확률을 $p[i]$ 라고 미리 알려져 있다는 가정과, i_1 번째 경주마가 1등, i_2 번째 경주마가 2등, ... , i_k 번째 경주마가 k 등 일 확률이

$$p_k[i_1, i_2, \cdots, i_k] = p_{k-1}[i_1, i_2, \cdots, i_{k-1}] \times \frac{p[i_k]}{1 - p[i_1] - p[i_2] - \cdots - p[i_{k-1}]}$$

임을 가정하고 입상확률을 유도하였다. 또한 Plackett (1975), Henery (1981) 는 각 경주마의 도착시간을 독립적인 지수분포 혹은 정규분포로 가정하여서

$$P(T_1 < T_2 < \cdots < T_n) = \int_{-\infty}^{\infty} f(t_1 - \theta_1) \cdots \int_{-\infty}^{\infty} f(t_n - \theta_n) dt_n \cdots dt_1$$

을 제안하였고, 이를 Stern(1990) 은 감마분포를 이용하여 적용하였다. 그러나, Harville 의 경우, 현실과 지나친 가정과 (Lo and Bacon-Shone, 1994), Henery 모형의 경우 다변량 적분을 거쳐야 하는 어려움이 있었다. 특히 두 모형은 모두, 각 경주마가 1등으로 안착할 확률을 그 경주에 배팅된 총 금액에 대한 그 경주마에 배팅 된 금액의 비율로 가정하였다. 그 후, Lo 와 Bacon-Shone(1994) 은 i-번째 경주마가 1등에 들어올 확률을

$$p_i[1] = F_i^{\beta} / \sum_r F_r^{\beta}$$

로 가정하고 (여기서 F_i는 i-번째 경주마에 배팅된 총액), 최우추정량을 사용하여 β를 구하는 방법을 보여주었다. 그러나, 이런 접근방법들은 모두 경주마의 우승확률을 추정하는데 말에 걸린 금액만을 고려한 것인데, 실제로 경마장에서 금액은 마감 전까지 시시각각으로 변하고 있으며, 경주마에 대한 산지, 성별, 나이, 말의 주행습성 등, 중요한 변수는 사용되고 있지 않았다. 이는 아마도 경마에 대한 분석보다는 이를 이용한 주식시장에의 활용을 기대한 이유일 가능성이 높아 보인다. 본 논문은 이들의 연구와는 달리, 일반 경마관람자가 경마장에 들어섰을 때 흔히 얻을 수 있는 기본적인 정보를 가지고 연승식에서 경주마가 당첨될 확률을 예측하는 모형을 로지스틱 모형을 통해 구해 보려고 한다.

2. 경마의 용어 및 진행

경마장에 들어서면, 전광판에 여러 종류의 경주가 계시되고 있는 것을 발견하게 된다. 각 경주마다 경주마에 대한 자료를 전광판이나 안내책자를 통해 쉽게 얻을 수 있다. 경주의 종류는 크게 장거리(2000 미터 이상), 단거리(1400 미터 이하), 중거리(1700-1900 미터)로 나뉘어 지고, 마권에 입상 예상마를 표시하여 매표소에 배팅금액을 지불함으로써 경마는 시작된다. 입상을 예상한 말이 적중될 확률을 '적중확률 혹은 입상확률'이라고 하며, 적중될 경우 그 말에 걸린 전체 금액에서 약 25%의 세금 및 수수료를 제외한 금액을 배팅금액비율로 나누어 환급 받게 된다. 각 경기 당 배팅하는 방법은 세 가지로 되어 있는데, '단승식(win)'이란 1등으로 들어오는 말을 적중시키는 방식이고, '연승식(place 혹은 show)'은 3등 이내 (3등 포함)로 들어오는 말을 적중시키는 방식이며, 마지막으로 '복승식(double)'은 1, 2 등을 순서에 관계없이 적중시키는 방식이다. 본 연구는 연승식으로 범위를 제한하고

있는데, 이는 로지스틱 모형의 기본가정인 이항분포가 독립성 가정을 내포하기 때문에, 연승식 경우 정확히 독립은 아니지만 단승식, 복승식에 비해 어느 정도 독립성을 유지할 것이라는 가정 때문이다.

경주마에 대한 자료는 한국마사회에서 나오는 그 날의 출마표가 나오는데, 여기에는 경주마의 마번, 마명, 산지, 연령, 성별, 부담중량, 기수명, 조교사명, 마주명, 나이, 데뷔일, 최근 진료내역, 최근 1년 전적 등이 나와 있다. 본 연구에서는 이 변수들 가운데 한국마사회에서 정기적으로 발간되는 '경마성적집'에 나타난 변수들을 고려하였다. 따라서, 산지(호주, 뉴질랜드, 한국), 연령, 성별(암, 수, 거세), 부담중량, 나이, 날씨, 노면상태를 고려하였고, 기수명과 조교사명은 숫자가 너무 많아 변수화 시키지 못하였다. 그 외에 경마예상지에서 쉽게 구할 수 있는 경주마의 주행특성(선행, 선입, 자유, 추입)도 고려해 보았다. 선행형이라는 것은 앞장서 무리 지어 달리다가 기회가 오면 선두로 나서려 하는 형이고, 선입형은 다른 말이 접근하면 점점 가속을 하는 특성이 있으며, 추입형은 후위에서 기회가 오면 추월하는 성격을 가지고 있으며, 자유형은 위의 습성을 모두 발휘하는 형이다. 본 연구에서는 1996년 10월 5일 제 71 경마부터 1996년 12월 1일 제 88 경마까지 총 205개의 경주 중에 취소, 실격이 포함되어 있는 경우를 제외한 경주자료를 분석에 사용했다.

3. 주생습성과 산지

경마장에서는 "단거리 경주는 선행형을 선택하라" 또는"장거리 경주는 자유형이 유리하다" (김문영, 1994) 라는 속설이 있다. 과연 경주마의 주행습성이 거리별 경주에 따라 다른 양상을 보이는지 알아보기 위해 로지스틱 회귀를 통해 분석해 보았다. 다음 표 3.1은 단거리, 중거리, 장거리에 있어서 주행습성에 따른 연승식의 입상률을 나타내고 있다.

표 3.1: 거리에 따른 경주별 주행습성 입상률

	단거리	중거리	장거리
선행형	24/51=0.470	26/55=0.472	18/45=0.400
선입형	12/51=0.235	8/50=0.160	13/49=0.265
자유형	6/51=0.117	4/55=0.072	6/45=0.133
추입형	0/24=0.000	0/25=0.000	2/21=0.095
χ_0^2	26.84	33.23	11.57
p-value	0.000	0.000	0.009

이 표에 의하면 각 거리별 경주에 대해 동질성 검정을 각각 실시한 결과, 주행습성에 따라 입상률이 다르다는 것을 알 수 있었다. 각 경주당, 선행형이 입상할 확률이 제일 높은 것으로 나타났는데, 선입형과의 차이도 각각 0.235, 0.312, 0.135이고 표준오차가 0.091, 0.084, 0.096이므로, 유의수준 0.05에서 장거리 경주를 제외하고는 모두 선행형이 선입형보다 우세하게 나타났다. 따라서, 단거리, 중거리 경주에서는 선행형이 연승식에 입상할 활률이 높

다고 볼 수 있다.

한편, 경주마의 산지별 연승식 입상분포를 살펴보면 표 3.2와 같게 된다.

표 3.2: 거리에 따른 경주별 산지 입상률

	단거리	중거리	장거리
호주산	19/73=0.260	13/70=0.185	19/48=0.395
한국산	3/14=0.214	8/26=0.307	4/15=0.266
뉴질랜드산	20/90=0.222	27/99=0.272	16/97=0.164
χ_0^2	0.367	2.285	9.333
p-value	1.832	0.319	0.009

여기에서, 동질성 검정의 카이제곱 통계량을 살펴보면, 장거리 경주에서만 산지별로 입상률이 다르다는 것을 알 수 있다. 또한 장거리 경주 경우, 호주산 경주마가 제일 높은 입상률을 나타내는데, 한국산 경주마의 입상률과의 차이가 유의하지 않고(유의확률 0.49), 한국산과 뉴질랜드산 입상률의 차이도 유의하지 않았다. (유의확률 0.2) 반면에 호주산과 뉴질랜드산 사이의 입상률 차이는 유의확률 0.0018 로 유의하였다. 따라서, 장거리 경주의 경우, 뉴질랜드산보다는 호주산을 선택하는 것이 유리할 것이다.

4. 나이와 부담중량

경주마의 나이가 적을수록 입상할 확률이 높을 것인가 하는 질문에 대답하기 위하여, 나이를 연속형 (3-9세) 독립변수로 간주하고, 로지스틱 회귀를 적합시켜서 그 우도비 통계량 (Likelihood Ratio Statistic) 의 로그를 취한 값을 이탈도(Deviance)라고 하며

$$G^2 = -2[L_M - L_S]$$

로 정의된다. 이때, L_M 은 모형의 로그우도 함수값이고 L_S 는 포화모형의 로그우도함수 값이 되며, 이 G^2 값은 적합 결여를 나타내는 카이제곱분포를 따르게 된다. (McCullagh and Nelder, 1989) 만일 모형의 부적합성을 나타내는 위 값이 허용치 보다 작게되면 로지스틱 회귀가 적합하다고 판단되고, 로지스틱 회귀의 나이변수에 대한 회귀계수 추정치를 살펴봄으로써, 나이와 연승식 적중률과의 관계를 추론할 수 있게 된다. 표 4.1은 나이와 연승식 적중률과의 로지스틱 회귀에 의한 값을 나타내며, 각 경주별 모형을 SAS 의 PROC GENMOD 로 적합한 내용이다. 따라서, 나이가 적은 경주마일 수록 단거리, 중거리, 장거리에서 모두 입상확률이 높게 된다.

이제, 부담중량과 입상확률에 대해 살펴보자. 부담중량이란 경마에서 우승의 기회를 균등히 하기 위해 마체에 부담을 주는 무게를 말하며 기수의 체중과 안장무게 등을 포함한다. 다시 말해서, 잘 뛰는 말은 무거운 납덩이를 등에 얹고, 못 뛰는 말은 기수의 체중을 줄이도록 하는 것이다. 수말이 암말보다 1kg 많고, 외국산 말이 국산 말보다 2kg 많게 된다. 또

표 4.1: 거리별 경주에 따른 경주마 나이와 연승식 입상률 관계

	단거리	중거리	장거리
Intercept	0.847(1.019)	4.213**(1.097)	0.799(0.861)
나이	-0.508**(0.258)	-1.068**(0.226)	-0.327**(0.146)
G^2	189.736	187.218	172.276
p-value	0.211	0.603	0.206

(** 는 유의수준 0.05 에서 유의한 값)

표 4.2: 거리별 경주에 따른 부담중량과 연승식 입상확률 관계

	단거리	중거리	장거리
Intercept	2.367(6.962)	-16.609**(6.109)	-11.118(4.839)
나이	-0.064**(0.127)	0.281**(0.110)	0.181**(0.087)
G^2	193.712	210.018	173.137
p-value	0.158	0.190	0.194

(**는 유의수준 0.05 에서 유의한 값)

한 수득상금과 우승횟수에 따라 부담중량이 늘어난다. 표 4.2는 부담중량과 연승식 입상확률에 대한 로지스틱 회귀분석의 결과이다. 단거리 경주의 경우, 부담중량이 유의한 영향을 끼치지 않지만, 중거리와 장거리 경주의 경우, 부담중량이 증가할 수록 오히려 입상확률이 증가하는 것을 알 수 있다. 이는 부담중량이 유명한 말일 수록 부담중량이 많게 되는 이유라고 추측된다.

5. 입상확률에 대한 예측모형

입상확률이란 어떤 경주마가 이 경주에서 입상할 수 있는 확률을 의미한다. 연승식 경우, 10마리 이상의 말 중에서 1, 2, 3등 내에 들어올 확률을 의미한다. 이 단원에서는 입상확률에 영향을 줄 수 있는 가능한 여러 가지 변수(주행습성, 성별, 나이, 부담중량, 노면상태(양호, 다습, 포화, 불량), 날씨, 출발위치 와 각각의 2차 교호작용까지) 들을 사용하여 stepwise 로 변수선택을 해 보았다. SAS 의 PROC LOGISTIC 이라는 프로시저에 stepwise 옵션을 사용한 결과, 다음의 결과를 얻었다. 여기서 π는 연승식에 입상할 확률을 의미하며, 모든 회귀계수는 유의수준 0.05에 유의한 것으로 판정되었다.

* 단거리:

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 0.508 - 1.297\ t1 + 0.429\ ts32 - 0.184\ tage2$$

여기서, $t1$은 선행형인 경우에 1, 선행형이 아닐 경우에는 -1 을 갖고, $ts32$는 경주마가 거세마이면서 주행습성이 자유형, 혹은 거세마도 자유형도 아닌 경우에는 1 을, 거세마이면서 자유형이 아니거나, 거세마 아니면서 자유형인 경우에는 -1 을 갖는다. $tage2$라는 변수는 선입형인 경우 나이를 나타내며, 선입형이 아닌 경우에는 -1 을 곱한 나이를 의미한다.

* 중거리:

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -3.233 + 0.673\ t3 + 0.942\ age + 0.3982\ rt21$$

변수 $t3$는 경주마가 자유형일 경우 1 을, 아닐 경우 -1 을 갖고, $rt21$은 노면상태가 다습(수분함량 10-15%)이고 선행형 경주마일 경우에는 1, 노면상태가 다습이 아니고 경주마가 선행형이 아닐 경우에는 -1 을 갖는 변수이다.

* 장거리:

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 0.769 - 0.094\ tage1 - 0.090\ hage2$$

여기서, $tage1$은 경주마가 선행형일 경우 경주마의 나이를 나타내며, 선행형이 아닐 경우에는 -1 을 곱한 나이를 의미한다. 한편, $hage2$는 호주산 경주마인 경우 경주마의 나이를 나타내고, 호주산이 아닐 때, -1 을 곱한 나이를 의미한다.

이와 같이 구한 연승식 입상확률 예측모형이 과연 어느 정도 신뢰성을 갖고 있는지 알아보기 위해, 1996년 10월 5일 부터 1996년 12월 1일 까지의 추정에 사용된 자료와는 별개의 172 경주를 가지고 평가를 해 보았다. 먼저, 각 경주에 참가하는 출주마의 정보를 가지고, 앞에서 구한 예측모형을 사용하여 각 경주마의 입상확률을 예측해 본다. 예측된 입상확률 $\hat{\pi}_i$ 가운데서 가장 확률이 높은 경주마를 선택하게 되며, 만일 두 마리의 말이 같은 확률로 값이 가장 크다면, 두 마리를 모두 선택하게 된다. 이렇게 선택된 말 중에서 실제 경기에서 연승식으로 입상된 경우를 %로 나타내면 이것이 예측모형의 성공률이 될 것이다. 다음 표 5.1은 단거리, 중거리, 장거리 경주에 대한 예측모형식의 성공률을 나타낸 것이다. 이는 단거리 예측모형의 경우 39%의 성공률을 갖고 있고, 중거리와 장거리는 각각 51%, 66%의 성공률을 갖게 된다. 이 수치는 12마리가 출주한다고 가정할 경우, 랜덤추출에 의한 자연성공률 25% 보다 유의수준 0.05에서 높다고 볼 수 있게 된다. 장거리 경주의 경우, 좀 더 많은 수의 자료가 있었으면 하는 아쉬움이 있지만, 같은 기간 내의 경주들을 사용하고자 했기 때문에 장거리 경주의 수가 상대적으로 단거리와 중거리에 비해 적게 사용되었다.

표 5.1: 로지스틱 회귀모형을 통한 예측모형의 성공률

	단거리	중거리	장거리
성공률	44/111=39%	28/54=51%	4/6=66%

6. 결론

앞에서 살펴본 바와 같이, 이 논문에서는 경마의 연승식의 입상률에 영향을 주는 변수들을 통계적으로 분석해 보았고, 입상률의 예측모형을 구현해 봄으로써, 관심 있는 변수들을 모형화 시켜 보았다. 선행연구들이 기존의 예측모형들을 경주마에 걸린 배팅금액을 위주로 이루어 졌지만, 본 연구에서는 경마장의 출마표에 근거한 경마 경주의 정보를 사용해 예측모형을 유도하였다. 따라서, 이 연구는 경마를 주식시장과 같은 내기시장으로 보지 않고, 경마장 내의 정보만을 활용해 분석을 시도해 보았다는 점에서 의미가 있을 것이다. 본 논문을 마치면서, 경주마에 대한 수상경력, 마주, 기수에 대한 정보 등을 고려할 수 없었던 점은 아쉬움으로 남으며, 추후 과제로 미루고 싶다. 또한 이 논문에 대한 조언을 아끼지 않으신 익명의 편집위원에게도 감사를 드리고 싶다.

참고문헌

[1] Ali, M. M. (1979). "Some Evidence on the Efficiency of a Speculative Market", Econometrica, 47, 387-392.

[2] Ali, M. M. (1994). "Probability Models on Horse Race Outcomes", ASA Proceedings of Section on Statistics in Sports, 7-11.

[3] Harville, David A. (1973). "Assigning Probabilities to the Outcomes of Multi-Entry Competitions", Journal of the American Statistical Association, Vol 68, No. 342, 312-316.

[4] Henery, R. J. (1981). "Permutation Probabilities as Models for Horse R aces", Journal of the Royal Statistical Society, Series B, Vol 43, 86-91.

[5] Lo, V. S. Y., Bacon-Shone, John (1994). "A comparison between two models for predicting ordering probabilities in multi-entry competition", The Statistician, Vol 43, No. 2, 317-327.

[6] McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, Chapman & Hall.

[7] Plackett, R. L. (1975). "The analysis of permutations", Applied Statistics, Vol 24. 193-202

[8] SAS Institute (1990). SAS/User's Guide, Version 6.

[9] Snyder, Wayne W. (1978). "Horse Racing: Testing the Efficient Markets Model" The Journal of Finance, Vol 33. No.4, 1109-1118.

[10] Stern, H. (1990). "Models for distributions on permutations", Journal of the American Statistical Association, Vol 85, 558-564.

[11] 김문영 (1994). 알기 쉬운 경마여행, 울도서적

[1998년 9월 접수, 2000년 2월 채택]

The Horse Race Winning Probability via Logistic Regression

Sun-Kyung Yoo[1] Heungsun Park [2]

ABSTRACT

The statistical inferences are made on the popular hypotheses concerned with the horse races in Korea and the winning probability is modeled with respect to many candidate variables from real data in Korea. This paper will be a good example for the statistics-related classes such as the logistic regression or decision making theory classes.

Keywords: Test of Homogeneity; Stepwise Selection; Contingency Table.

1) Graduate Student, Dept. of Statistics, Seoul National University.
2) Associate Professor, Dept. of Statistics, Hankuk University of Foreign Studies.
E-mail: hspark@stat.hufs.ac.kr