The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.2.223
26(2), 223–236

Implementation of Mahalanobis-Taguchi System for the Election of Major League Baseball Hitters to the Hall of Fame

Su Whan Kim[a] · Changsoon Park[a,1]

[a]Department of Statistics, Chung-Ang University

(Received December 31, 2012; Revised February 6, 2013; Accepted February 7, 2013)

Abstract

Various statistical classification methods to predict election to the Major League Baseball hall of fame of are implemented and their accuracies are compared. Seventeen independent variables are selected from the data of candidates eligible for the hall of fame and well-known classification methods such as discriminant analysis and logistic regression as well as the recently proposed Mahalanobis-Taguchi system(MTS). The MTS showed a better performance than the others in classification accuracy because it is especially efficient in cases where multivariate data does not constitute directionally geographical groups according to attributes.

Keywords: Mahalanobis-Taguchi System, discriminant analysis, logistic regression, classification accuracy.

1. 서론

미국 프로야구(Major League Baseball)는 1869년 첫 프로구단이 창단한 이후 1901년에 지금과 같은 아메리칸 리그와 내셔널 리그의 양대 리그 체제를 구축하였고 현재는 30개 팀으로 구성되어 있다. 140여년이 넘는 역사를 가진 메이저리그는 최고의 실력을 가진 수많은 선수들이 거쳐 간 세계 최고의 야구리그로서 현재까지 293명의 선수만이 명예의 전당에 입성하였으며 비록 명예의 전당에 입성하지 못했지만 뛰어난 실력을 가진 수많은 선수들의 기록이 남아 있는 역사와 전통을 자랑하는 리그이다.

명예의 전당은 야구탄생 100주년과 뛰어난 활약을 보인 선수들을 기념하기 위해 1939년에 설립되었고 1936년부터 입성을 시작했다. 명예의 전당에 입성하는 방법으로는 미야구기자단협회(Baseball Writers Association of America)와 원로 위원회(Veterans Committee), 니그로리그 위원회(Negro League Committee) 등을 통하여 입성하는 방법이 있다. 일반적으로 많은 선수들이 미야구기자단협회의 투표를 통하여 명예의 전당에 입성을 하게 된다. 이를 위한 조건은 메이저리그에서 선수경력이 10년 이상되고 은퇴 후 5년의 유예기간이 지나야 후보자격이 생기며 이후 15년까지 기회가 주어진다. 미야구기자협회의 투표에서 75%이상의 찬성 득표율을 얻어야 명예의 전당에 입성을 하고 찬성득표율이 5%이하이면 남은 기회와 상관없이 후보자격을 상실한다. 은퇴 후 5년이 지나서 자격이 주어지는 이유는 은퇴

This research was supported by the Chung-Ang University Research Scholarship Grants in 2011.
[1]Corresponding author: Professor, Department of statistics, Chung-Ang University, Seoul 156-756, Korea. E-mail: cspark@cau.ac.kr

후 바로 투표를 하면 투표권자가 특정 선수의 은퇴 무렵에 일어나는 과열된 분위기에 휩싸이기 때문에 객관적인 판단을 할 수 없기 때문이다. 미야구기자협회의 투표에서 15번의 기회가 될 때까지 찬성률이 75%를 넘기지 못한 선수는 원로 위원회의 심사대상이 되며 원로위원회는 선수출신 원로인, 야구행정관계자, 원로기자 등 15인으로 구성되며 만장일치가 되면 명예의 전당에 입성이 된다. 지금은 해체 되었지만 1971년부터 1977년까지 니그로리그 위원회가 활동하여 니그로리그 선수들이 명예의 전당에 입성했고 그 당시 시대 상황에 따라 명예의 전당에 입성하는 기준이 임시적으로 생기는 경우가 있으나 명예의 전당 입성자수는 매우 적었다. 또한 선수뿐만 아니라 야구발전에 공헌한 감독, 구단주, 기자 등도 명예의 전당에 입성할 수 있으며 이러한 입성자들은 각자의 분야에서 인정과 존경을 받고 야구발전에 공헌한 사람들이다 (Hample, 2007; James, 2001).

통계적인 관점에서 야구를 분석하는 방법론인 세이버매트릭스(sabermatrics)에서는 명예의 전당의 입성가능성에 대하여 측정하는 지표로서 4가지(Black Ink Test, Grey Ink Test, HOF Monitor, HOF Career Standards)를 사용하고 있다. 이 지표들은 일반적으로 명예의 전당의 입성과 탈락 여부에 대하여 기존 입성자의 평균 점수와 비교하여 가능성을 판단하지만 입성 가능성에 대한 단순 비교일 뿐 예측하는 것이 아니다. 세이버메트릭스에 대한 4가지 지표는 “http://www.baseball-reference.com/about/leader_glossary.shtml”에서 자세히 설명되고 있다.

이 연구에서는 명예의 전당 입성조건을 갖춘 선수들을 대상으로하여 입성과 탈락으로 분류할 때, 통계적인 널리 사용되어온 분류방법인 판별분석, 로지스틱 회귀분석과 비교적 최근에 제안된 마할라노비스-다구찌 시스템(Mahalanobis-Taguchi System; MTS)을 사용하여 분석하고 비교하였다. 판별분석과 로지스틱 회귀분석은 Kwon (2008a), Lee (2011), Huh와 Yang (2001) 등에서 설명되고 있으며 MTS는 Taguchi 등 (2005), Taguchi와 Jugulum (2000, 2002)에 나타나 있다. 특히 MTS를 포함한 이유는 아무리 우수한 선수도 모든 항목이 우수할 수 없고 일부 항목에서는 평균보다 못한 항목들이 있어 다변량 자료가 입성의 경우 특히 방향성이 없이(direction-invariant) 도형적인(geometrical) 그룹을 형성하지 못하는 경우가 있다. 이와 같이 두 그룹 분류분석에서 한 그룹이 하나의 도형적인 그룹을 형성하는 반면에 다른 그룹은 하나의 도형적 그룹을 형성하지 못하는 경우에 MTS는 특히 효율적이기 때문이다 (Park, 2012).

2. 데이터분석 및 산점도

데이터의 수집은 베이스볼 레퍼런스(http://www.baseball-reference.com)에서 1980년 이후 명예의 전당 입성조건을 만족하는 255명의 선수들의 정보와 데이터를 이용했다. 1980년 이후를 기준으로 한 이유는 명예의 전당의 입성기준이 시대에 따라 투표방식이 조금씩 변했는데 1980년 이후의 선수들이 현재의 입성기준과 같이 적용된 선수들이고 현대야구의 투수분업화가 시작되는 지점으로 좀 더 체계적으로 데이터가 측정되었기 때문이다.

총 255명의 타자들 중 원로위원회를 통해 입성한 23명의 선수집단을 탈락집단에 포함시켜 분석을 실시하였다. 원로 위원회를 통해 입성한 선수들 역시 15번의 미기자단협회의 투표에서 찬성률 75%를 만족하지 못하고 자격을 상실하여 탈락집단과 대등한 것으로 간주한다. 미기자단협회의 투표를 통한 36명의 명예의 전당 입성집단과 탈락자 및 원로위원회를 통한 명예의 전당 입성자로 구성된 219명의 탈락집단으로 구성하여 분석을 실시하였다.

명예의 전당의 입성과 탈락의 영향을 주는 요인으로 선정된 변수는 다음 17개의 변수는 Table 2.1에 나타나 있다.

Table 2.1의 변수들은 타자에 대한 기본적인 능력을 알 수 있는 BA, OBP, SLG, OPS과 같은 비율 기

Table 2.1. Variables affecting the election to Hall of Fame

변수	변수
시즌수(Seasons)	2루타(Double hit; 2B)
올스타게임 선정횟수(All-Star Games; ASG)	3루타(Triple hit; 3B)
게임수(Game played; G)	홈런(Home Runs; HR)
타석(Plate Appearances; PA)	타점(Runs Batted In; RBI)
타수(At Bats; AB)	도루(Stolen Base; SB)
득점(Run scored; R)	볼넷(Bases on balls; BB)
안타(Hits; H)	타율(Batting Average; BA)
출루율(On Base Percentage; OBP)	장타율(Slugging Percentage; SLG)
OPS(On Base Plus Slugging Percentage)	

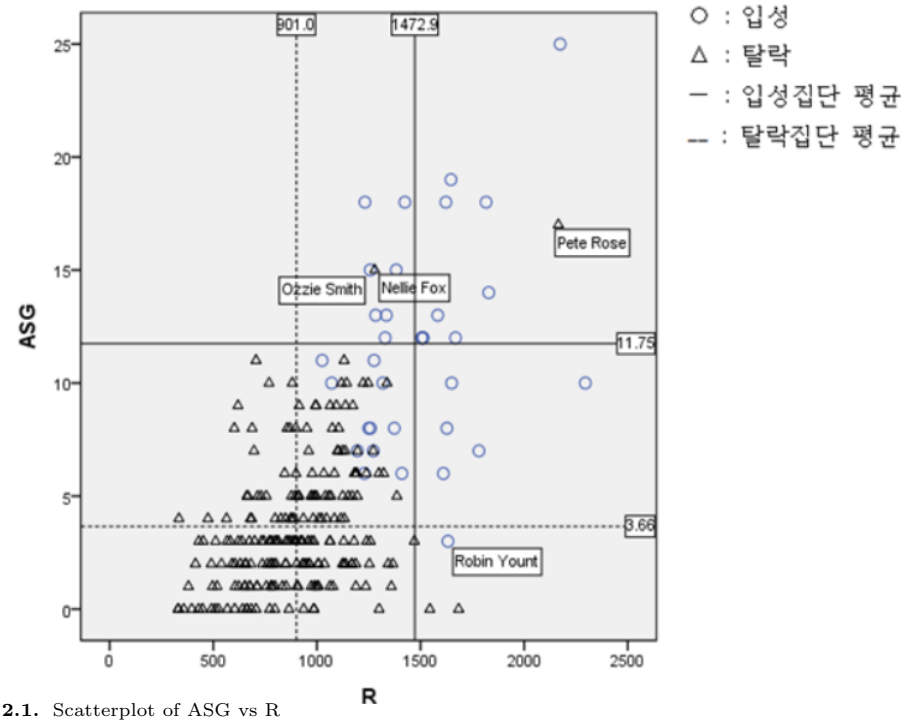

Figure 2.1. Scatterplot of ASG vs R

록과 G, PA, AB, R, H, 2B, 3B, HR, RBI, R, SB, BB과 같은 누적기록으로 구성된 변수들이며 이 변수들 중 ASG는 팬들의 인기투표로 선정되는 것으로서 실력과 더불어 인기도를 나타내고 시즌수는 뛰어난 실력을 얼마나 오랜 기간 유지했는지를 알 수 있는 변수이다 (James, 2001; Hample, 2007; Marsh, 2007).

먼저 17개의 변수 중 서로 상관관계가 있는 두 변수를 선택하여 데이터의 도형적 그룹이 형성되는 경우와 그렇지 못한 경우를 알아보았다.

Figure 2.1은 ASG와 R에 대한 산점도를 나타내고 있다. 이 그림에서 두 변수는 강한 양의 상관관계를 나타내는 것을 알 수 있으며 탈락자들은 뚜렷한 도형적 그룹을 형성하지만 입성자들은 상대적으로 넓게 산포되어 그렇지 못함을 알 수 있다. 명예의 전당 입성에 실패한 선수인 Pete Rose선수는 명예의 전당

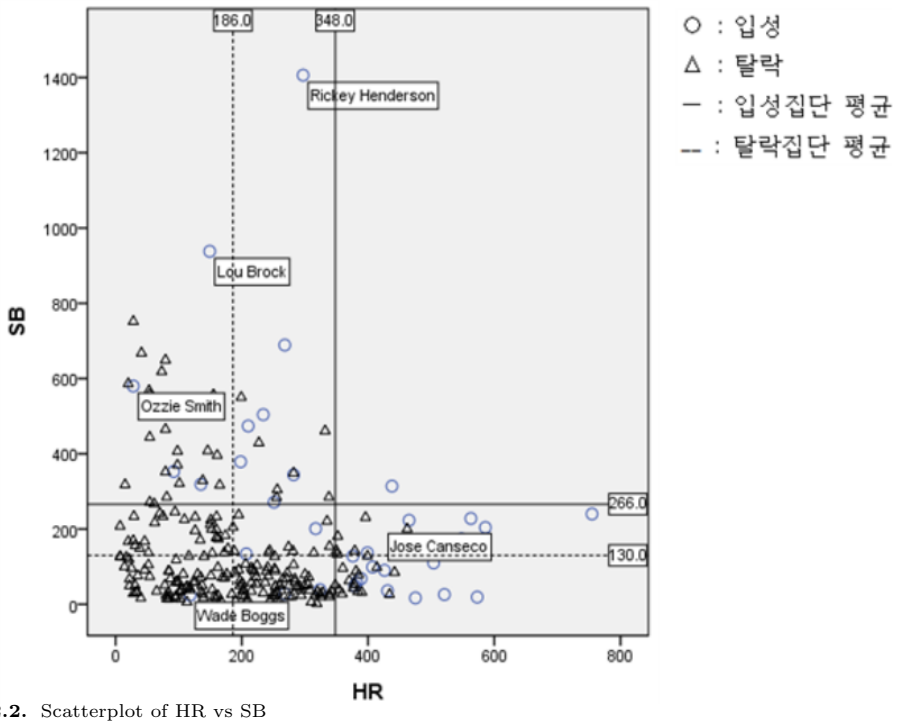

Figure 2.2. Scatterplot of HR vs SB

에 입성한 집단의 평균보다 더 높은 곳에 위치해 있으며 통산최다안타, 최다경기, 최다타수 1위를 기록한 선수로서 명예의 전당 입성이 확실시 되는 산점도상에 위치를 나타났지만 야구 승부도박과 탈세의 불명예스런 행위로 영구 제명되어 명예의 전당에 입성하지 못했다. Robin Yount선수는 20년 동안 한 팀에서만 뛰며 뛰어난 누적기록을 달성했지만 소속팀이 비인기 팀으로 올스타게임 출전에는 불리하게 작용했다. Ozzie Smith와 Nellie Fox선수는 산점도상에서 가장 근접하게 위치하였으나 Ozzie Smith선수는 수비부담이 큰 유격수로서 화려한 수비를 보인 선수로 누적기록뿐만 아니라 수비적인 측면도 인정받아 명예의 전당에 입성하게 되었다. Nellie Fox선수도 뒤지지 않는 실력을 보였지만 마지막 15번째 기회에서 명예의 전당 입성인 75% 찬성률에 0.3% 모자라는 74.7%를 기록하여 아쉽게 탈락하였다. 후에 원로위원회를 통해 명예의 전당에 입성하였지만 이 논문에서는 원로위원회를 통해 명예의 전당에 입성한 선수들은 탈락집단에 포함하는 것으로 간주했다.

Figure 2.2는 HR과 SB에 대한 산점도를 나타내고 있다. 이 그림에서 두 변수는 음의 상관관계를 나타내며 탈락자들은 비교적 양호한 도형적 그룹을 형성하지만 입성자들은 그렇지 못함을 알 수 있다. Ozzie Smith선수는 위에 언급되었고 Rickey Henderson, Lou Brock선수는 메이저리그를 대표하는 1번 타자로 통산도루를 1, 2위를 기록하는 선수들이다. 야구에서 1번 타자의 역할은 홈런보다는 출루와 도루를 통해 기회를 만들어내는 역할로서 그 가치를 인정받았다고 할 수 있다. Jose Canseco선수는 많은 홈런 수를 기록했지만 홈런능력에 비해 정교함이 떨어졌고 출전경기의 2/3가량을 지명타자로 경기를 나서 반쪽선수라는 평가를 받아 명예의 전당에 탈락 하였다. Wade Boggs선수는 홈런보다는 정확한 타격을 통해 안타를 많이 생산해내는 타자로 통산 3010안타를 기록하며 적은 홈런을 만회하는 실력을 보여준 선수이다. 메이저리그에서는 3000안타 이상이나 500홈런 이상을 기록하면 명예의 전당 입성의

Table 3.1. Multi-collinearity statistic according the independent variable

모형	회귀계수	유의확률	공선성 통계량	
			공차	VIF
(상수)	−0.791	0.373		
Seasons	−0.014	0.197	0.191	5.238
ASG	−0.036	0	0.479	2.088
G	8.15E-05	0.722	0.019	51.514*
PA	0	0.508	0	2954.628*
AB	0.001	0.174	0	3002.978*
R	0	0.206	0.025	39.255*
H	−0.001	0.039	0.003	312.041*
2B	−0.001	0.012	0.084	11.911*
3B	−0.002	0.023	0.248	4.03
HR	−0.003	0	0.027	37.302*
SB	0	0.091	0.046	21.705*
RBI	−0.001	0	0.339	2.954
BB	0	0.667	0.017	60.324*
BA	−1.707	0.626	0.049	20.332*
OBP	−0.398	0.909	0.028	36.194*
OPS	3.667	0.005	0.025	39.61*

매직넘버라고 말하는데 3000안타 이상을 기록한 선수들 중 위에서 언급한 Pete Rose를 제외한 모든 선수가 명예의 전당에 입성했으며 500홈런 이상 기록한 선수들 역시 Figure 2.2에서 보이는 것과 같이 모두 명예의 전당에 입성하였다.

3. 다중공선성과 타당성

3.1. 다중공선성

이 논문에서는 명예의 전당 입성조건을 만족하는 타자들에 대한 데이터의 독립변수들에 대한 다중공선성진단을 하였다. 이때 VIF(Variation index factor)의 값이 클 경우 VIF의 변수가 다른 독립변수들에 의해 선형함수를 표현될 수 있게 되어 이로 인해 다중공선성 문제가 발생한다. 일반적으로 VIF값이 10이상인 독립변수가 다중공선성 문제를 발생시킨다고 판단한다 (Kwon, 2008b).

Table 3.1은 데이터 전체인 총 17개의 독립변수에 대한 VIF의 값을 보여주고 있다. 이때 G, PA, AB, R, H, 2B, HR, SB, BB, BA, OBP, OPS 등 12개의 독립변수가 10이상의 VIF값을 나타내어 다중공선성 문제를 발생시킨다고 판단된다.

Table 3.1에서 VIF가 10미만인 변수는 Seasons, ASG, 3B, 그리고 RBI 등 4개밖에 되지 않아, 변수의 수를 증가시킬 필요가 있다고 판단되었다. 변수의 수를 증가시키기 위해 Table 3.1의 VIF가 10이상인 변수들 중 상대적으로 적은 VIF값을 가지는 변수들을 대상으로 모형에 포함시켰을 때 VIF가 10미만이 되는 변수들을 찾아내어 그 결과를 Table 3.2에 나타내었다. 이들 변수그룹에는 Seasons, ASG, 2B, 3B, HR, RBI, BB, BA, OPS 등 총 9개의 독립변수가 포함되었으며 다중공선성 문제를 유발하지 않는 것으로 판단되었다. 이들 변수그룹을 공선성고려 변수라 한다.

전체 독립변수를 이용한 방법과 위의 다중공선성진단을 통하여 다중공선성이 의심되지 않는 독립변수를 이용한 방법으로 각 분류 분석방법에 적용하였다.

Table 3.2. Variables not inducing multi-collinearity

모형	회귀계수	유의확률	공선성 통계량	
			공차	VIF
(상수)	1.435	0		
Seasons	−0.009	0.263	0.415	2.411
ASG	−0.037	0	0.531	1.884
2B	6.28E-05	0.82	0.251	3.991
3B	6.55E-05	0.923	0.512	1.952
HR	−0.001	0.003	0.138	7.228
RBI	0	0	0.649	1.541
BB	−8.25E-05	0.312	0.327	3.058
BA	−2.475	0.166	0.203	4.937

3.2. 타당성(validity)

분류 방법의 객관적인 평가를 위해 예비법(holdout method)과 이중교차법(Two-fold cross validation)을 사용하였다. 예비법에서는 255명의 전체 데이터 중 서로 겹치지 않게 128명의 훈련용 데이터(training data)와 나머지 127명을 시험용 데이터(test data)로 랜덤하게 분할한 다음, 훈련용 데이터로 분류기준을 수립하고 시험용 데이터를 이용하여 분류의 정확도를 계산하는데 이 과정을 100번 반복하여 평균 분류정확도를 계산하였다. 이중 교차타당성 방법에서는 위의 예비법과 동일한 과정을 거치지만 훈련용과 시험용의 역할을 바꿔가며 실시한다.

4. 분류방법

메이저리그 명예의 전당 입성조건을 만족하는 255명의 타자들을 대상으로 명예의 전당의 입성과 탈락의 결과를 예측하기 위하여 판별분석, 로지스틱 회귀분석, MTS방법을 적용하여 그 결과를 비교하였다.

4.1. 판별분석

명예의 전당 입성 자격을 갖춘 255명의 선수들을 명예의 전당 입성과 탈락으로 분류하고 판별분석을 실시하여 판별할 기준과 규칙을 만들어 명예의 전당의 입성집단과 탈락집단이 얼마나 잘 분류 되는가를 알아보고자 한다. 판별분석은 집단 간의 차이를 판별해주는 유용한 방법 중 하나로서 사전확률은 균일분포로 설정하였다. 앞 절에서 설명한 17개의 변수를 이용하여 명예의 전당 입성과 탈락 여부인 종속변수의 집단구분을 예측하고자 한다.

Table 4.1은 명예의 전당의 입성집단과 탈락집단의 평균과 표준편차가 제시되어 있다. 전체적으로 명예의 전당에 입성한 집단의 평균이 탈락한 집단에 비해 상대적으로 높게 나타나는 것을 확인할 수 있다. 특히 누적기록에 있어서 오랜 기간 선수생활을 하는 것이 유리하며 이는 Seasons의 평균차이가 전체적인 누적기록의 차이를 나타내는 이유라고 볼 수 있다.

Table 4.2는 전체 독립변수를 이용하여 예비법을 수행한 판별분석결과이며 시험용 데이터의 분류정확도는 입성집단 67%, 탈락집단 98%, 전체 94%로 나타났다.

Table 4.3은 공선성고려 독립변수를 이용하여 예비법을 수행한 판별분석결과이며 시험용 데이터의 분류정확도는 입성집단 69%, 탈락집단 98%, 전체 94%로 나타났다.

Table 4.4는 전체 독립변수를 이용하여 이중 교차타당성 방법을 수행한 판별분석결과이며, 분류정확도는 입성집단 67%, 탈락집단 98%, 전체 94%로 나타났다.

Table 4.1. Statistics for election and elimination groups to Hall of Fame

변수	입성 ($n_1 = 36$)		탈락 ($n_2 = 219$)	
	평균	표준편차	평균	표준편차
Seasons	20.056	2.64	15.973	2.704
ASG	11.75	4.594	3.662	2.982
G	2595.75	342.637	1858.772	389.204
PA	10838.58	1552.193	7318.137	1756.718
AB	9531.694	1358.71	6500.653	1562.367
R	1472.917	281.241	901.151	265.633
H	2735.278	440.785	1792.265	468.975
2B	477.389	91.736	305.438	92.841
3B	74.75	29.121	48.772	30.07
HR	348.056	166.213	186.046	110.97
SB	1394.306	343.773	869.539	303.926
RBI	266.028	288.543	130.37	141.692
BB	1115.972	374.535	671.676	270.761
BA	0.287	0.02	0.275	0.018
OBP	0.362	0.024	0.344	0.025
SLG	0.464	0.049	0.422	0.053
OPS	0.825	0.062	0.766	0.069

Table 4.2. Discrimination analysis using whole independent variables (Holdout method)

	입성	탈락	분류정확도(%)
입성	11.88	5.97	67
탈락	1.9	107.25	98
전체			94

Table 4.3. Discrimination analysis using variables not inducing multi-collinearity (Holdout method)

	입성	탈락	분류정확도(%)
입성	12.38	5.58	69
탈락	1.77	107.27	98
전체			94

Table 4.4. Discrimination analysis using whole independent variables (two-fold cross validation)

	입성	탈락	분류정확도(%)
입성	11.99	5.58	67
탈락	2.06	107.10	98
전체			94

Table 4.5. Discrimination analysis using variables not inducing multi-collinearity (two-fold cross validation)

	입성	탈락	분류정확도(%)
입성	12.04	5.53	69
탈락	1.81	107.62	98
전체			94

Table 4.5는 다중공선성 진단을 통해 선정된 독립변수를 이용하여 이중 교차타당성 방법을 수행한 판별분석결과이며, 분류정확도는 입성집단 69%, 탈락집단 98%, 전체 94%로 나타났다.

Table 4.6. Estimated regression coefficients in logistic regression

변수	B	유의확률	변수	B	유의확률
Seasons	−0.057	0.655	HR	−0.012	0.161
ASG	−0.144	0.007	SB	0.001	0.575
G	0	0.907	RBI	−0.002	0.192
PA	−0.001	0.817	BB	−0.001	0.879
AB	0.002	0.637	BA	−6.816	0.881
R	0.001	0.671	OBP	11.273	0.972
H	−0.004	0.484	SLG	13.041	0.968
2B	−0.005	0.386	OPS	1.632	0.996
3B	−0.009	0.421	(상수)	−5.113	0.649

Table 4.7. Logistic regression using whole independent variables (Holdout method, $T_L = 0.69$)

	입성	탈락	분류정확도(%)
입성	17.34	0.70	96
탈락	5.44	104.51	95
전체			95

Table 4.8. Logistic regression using variables not inducing multi-collinearity (Holdout method, $T_L = 0.67$)

	입성	탈락	분류정확도(%)
입성	17.13	1.00	94
탈락	5.29	104.58	95
전체			95

판별분석의 결과를 종합하면 전체변수나 공선성고려 변수에 관계없이 분류정확도는 아주 근접하게 나타났으며 에비법이라 이중교차법 간에도 아무런 차이점을 발견할 수 없었다. 전반적으로 94%대의 양호한 분류정확도를 나타내고 있음을 알 수 있다.

4.2. 로지스틱 회귀분석

명예의 전당 입성 자격을 갖춘 255명의 선수들을 대상으로 명예의 전당 입성(0)과 탈락(1)으로 구분하여 로지스틱 회귀분석을 실시하였으며 그 분류결과의 정확성에 대해를 알아보았다. 이 때 로지스틱 회귀분석에서는 분류 분리점(threshold, T_L)을 설정할 때 두 종류의 오분류(입성일 때 탈락으로 분류, 탈락일 때 입성으로 분류) 상대빈도를 동일하게 할 수 있도록 설정하였다. 17개의 설명변수에 대한 로지스틱 회귀의 결과는 Table 4.6과 같다.

Table 4.7은 전체 독립변수를 이용하여 예비법을 수행한 로지스틱 회귀분석결과($T_L = 0.69$)이며, 시험용 데이터의 분류정확도는 입성집단 96%, 탈락집단 95%, 전체 95%로 나타났다.

Table 4.8은 공선성고려 변수를 이용하여 예비법을 수행한 로지스틱 회귀분석결과($T_L = 0.67$)이고 시험용 데이터의 분류정확도는 입성집단 94%, 탈락집단 95%, 전체 95%로 나타났다.

Table 4.9는 전체 독립변수를 이용하여 이중 교차법을 수행한 로지스틱 회귀분석결과($T_L = 0.74$)이고 분류정확도는 입성집단 87%, 탈락집단 86%, 전체 86%로 나타났다.

Table 4.10은 공선성고려 변수를 이용하여 이중 교차법을 수행한 로지스틱 회귀분석결과($T_L = 0.74$)이고 분류정확도는 입성집단 89%, 탈락집단 90%, 전체 90%로 나타났다.

Table 4.9. Logistic regression using whole independent variables (two-fold cross validation, $T_L = 0.74$)

	입성	탈락	분류정확도(%)
입성	15.90	2.41	87
탈락	15.22	94.29	86
전체			86

Table 4.10. Logistic regression using variables not inducing multi-collinearity (two-fold cross validation, $T_L = 0.74$)

	입성	탈락	분류정확도(%)
입성	16.05	1.96	89
탈락	11.05	98.46	90
전체			90

로지스틱 회귀분석의 결과를 종합하면, 예비법에서는 대상변수에 관계없이 분류정확도가 동일하게 나타났으나 이중교차법에서는 전체변수를 사용한 경우의 정확도가 공선성고려 변수를 사용한 경우보다 조금 떨어지는 것으로 나타났다. 이 결과의 의미는 여타 통계적 방법에서와 같이 변수의 선택에 따른 장단점이라 할 수 있다. 즉, 변수를 선별하면 변수의 수가 줄어들어 정보수집과 분석이 용이하나, 정보의 손실에 따른 효율의 감소가 발생한다. 결과적으로 이 논문에서는 다중공선성 문제가 고려될 필요는 없었으나 그것이 분류분석에 어떤 영향을 미치는지에 대해서는 논의에서 제외한다.

4.3. MTS방법

마할라노비스 거리(Mahalanobis distance)는 인도의 수학자 Mahalanobis에 의해 한 집단에서 이질의 집단을 구분하는 방법으로 1930년대에 소개 되었다. 강건설계방법(robust design)을 고안해낸 다구찌는 어떤 집단의 평균값을 기초로 한 마할라노비스 공간(Mahalanobis space)을 설정하고 이를 기초로 한 새로운 관측값이 공간으로부터 얼마나 벗어나 있는가를 측정하는 MTS방법을 고안하였다(Taguchi와 Jugulum, 2000).

마할라노비스 거리는 변수들 사이의 표준편차와 상관관계를 고려되어 만들어진 거리로서 다음과 같이 나타낸다.

$$d = (X - \mu)' \Sigma^{-1} (X - \mu).$$

단, $X = (x_1, x_2, \ldots, x_p)$, μ는 평균, Σ는 공분산이다.

MTS방법은 전체 데이터를 정상그룹(N)과 비정상그룹(A)로 분류하고 정상그룹으로부터 마할라노비스 공간을 정의한 다음 각 개체가 그 공간의 중심으로부터 얼마나 떨어져있는가를 마할라노비스 거리로 나타내어 분류에 적용한다. MTS방법의 절차와 정상그룹과 비정상그룹의 데이터는 다음과 같이 표현한다.

$$X^N = \begin{pmatrix} x_{11}^N & x_{12}^N & \cdots & x_{1p}^N \\ x_{21}^N & x_{22}^N & \cdots & x_{2p}^N \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_N 1}^N & x_{n_N 2}^N & \cdots & x_{n_N p}^N \end{pmatrix}_{n_N \times p}, \qquad X^A = \begin{pmatrix} x_{11}^A & x_{12}^A & \cdots & x_{1p}^A \\ x_{21}^A & x_{22}^A & \cdots & x_{2p}^A \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_N 1}^A & x_{n_N 2}^A & \cdots & x_{n_N p}^A \end{pmatrix}_{n_A \times p}.$$

단, X^N은 정상그룹, X^A는 비정상 그룹, 전체데이터는 $n = n_N + n_A$이다.

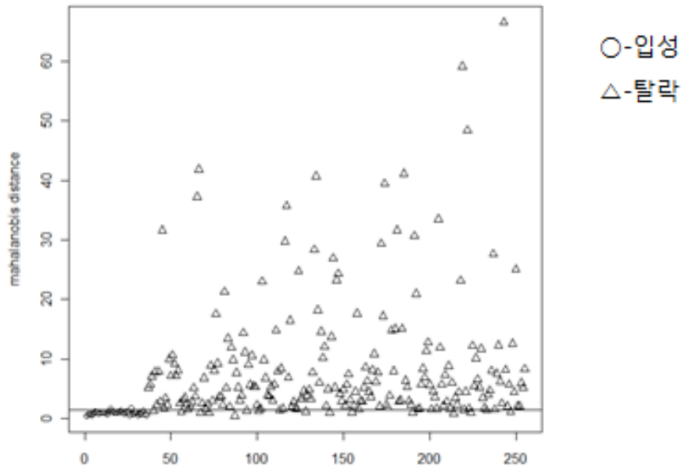

Figure 4.1. Scatterplot of MTS results

Table 4.11. MTS using whole independent variables (Holdout method, $T_M = 1.3$)

	입성	탈락	분류정확도(%)
입성	17.96	0.00	99
탈락	0.12	109.92	99
전체			99

Table 4.12. MTS using variables not inducing multi-collinearity (Holdout method, $T_M = 1.36$)

	입성	탈락	분류정확도(%)
입성	16.36	1.68	91
탈락	11.44	98.52	90
전체			90

정상그룹의 평균벡터를 $\mu^N = (\mu_1^N, \mu_2^N, \ldots, \mu_p^N)'$, 공분산행렬을 Σ^N이라 하면 각 데이터에 대한 마할라노비스 거리는 다음과 같이 정의한다.

$$d_i^N = \left(X_i^N - \mu^N\right)\left(\Sigma^N\right)^{-1}\left(X_i^N - \mu^N\right)', \quad i = 1, 2, \ldots, n_N,$$
$$d_j^A = \left(X_j^A - \mu^N\right)\left(\Sigma^N\right)^{-1}\left(X_j^A - \mu^N\right)', \quad j = 1, 2, \ldots, n_A.$$

이때 마할라노비스 거리가 분류분리점(T_M)보다 작으면 정상그룹으로, 크거나 같으면 비정상그룹으로 분류한다. MTS에서도 로지스틱 회귀분석에서 처럼 분류분리점을 설정할 때 두 종류의 오분류(입성일 때 탈락으로 분류, 탈락일 때 입성으로 분류) 상대빈도를 동일하게 할 수 있도록 설정하였다.

Figure 4.1은 전제 255명 타자들의 마할라노비스 거리를 산점도로 나타낸 것으로서, 왼쪽의 "◯"로 표시된 그룹은 입성집단, 오른쪽의 "△"로 표시된 그룹은 탈락집단을 나타낸다. 분류분리점은 수평선으로 나타나있다.

Table 4.11은 전체 독립변수를 이용하여 예비법를 수행한 MTS결과($T_M = 1.3$)이며 시험용 데이터의 분류정확도는 입성집단 99%, 탈락집단 99%, 전체 99%로 나타났다.

Table 4.12는 공선성고려 변수를 이용하여 예비법을 수행한 MTS결과($T_M = 1.36$)이고 분류정확도는

Table 4.13. MTS using whole independent variables (two-fold cross validation, $T_M = 1.45$)

	입성	탈락	분류정확도(%)
입성	17.96	0.05	99
탈락	0.30	109.20	99
전체			99

Table 4.14. MTS using variables not inducing multi-collinearity (two-fold cross validation, $T_M = 1.34$)

	입성	탈락	분류정확도(%)
입성	16.09	1.92	89
탈락	11.30	98.19	90
전체			89

Table 4.15. Results of MTS using the Mahalanobis space based on the elimination group

	입성	탈락	합	분류정확도(%)
입성	19	17	36	47
탈락	4	215	219	98
전체			255	91

입성집단 91%, 탈락집단 90%, 전체 90%로 나타났다.

Table 4.13은 전체 독립변수를 이용하여 이중교차법을 수행한 MTS결과($T_M = 1.45$)이고 분류정확도는 입성집단 99%, 탈락집단 99%, 전체 99%로 나타났다.

Table 4.14는 공선성고려 변수를 이용하여 이중교차법을 수행한 MTS결과($T_M = 1.34$)이고 분류정확도는 입성집단 89%, 탈락집단 90%, 전체 89%로 나타났다.

MTS 방법에서는 타당성방법(예비법, 이중교차법)의 종류에는 무관하게 분류정확도가 일정하게 나타남을 알 수 있다. 반면에 대상변수(전체변수, 공선성고려 변수)의 종류에 대해서는 분류정확도가 크게 차이가 남을 알 수 있다. 특히 전체 독립변수를 사용한 경우에는 매우 정확한 분류가 이루어지고 있음을 알 수 있다.

마할라노비스 공간을 정의할 때는 어떤 그룹을 정상그룹으로 정의하는가에 따라서 분류정확도의 차이가 날 수 있다. 이에 대한 효과를 알아보기 위해 탈락집단으로 마할라노비스 공간을 정의하여 보았다. Table 4.15은 마할라노비스 공간을 탈락집단으로 정의한 결과 ($T_M = 2.6$)이고 분류정확도는 입성집단 47%, 탈락집단 98%, 전체 91%으로 나타났다. Figure 4.2는 탈락집단을 마할라노비스 공간으로 설정한 후 계산한 마할라노비스 거리를 산점도로 나타낸 것으로서, 왼쪽의 "◯"로 표시된 그룹은 탈락집단, 오른쪽의 "△"로 표시된 그룹은 입성집단을 나타낸다.

이와 같이 탈락집단으로 마할라노비스 공간을 정의하면 입성집단을 대상으로 한 경우보다 분류정확도가 크게 떨어지는 것을 알 수 있다. 이 결과는 마할라노비스 공간을 정의할 때 어느 그룹을 대상으로 하느냐가 분류 효율에 영향을 줄 수 있음을 나타내고 있다. 하지만 이 문제는 두 그룹 각각에 대해 마할라노비스 공간을 정의하고 그 분류효과를 보면 어느 그룹으로 공간을 정의해야할지가 분명해진다.

5. 결론

이 논문에서는 명예의 전당 입성조건을 만족하는 타자들의 기록인 17개의 변수를 사용하여 판별분석, 로지스틱 회귀분석, MTS방법을 실시했다. 기존의 방법인 세이버 매트릭스의 명예의 전당 지수는 단순

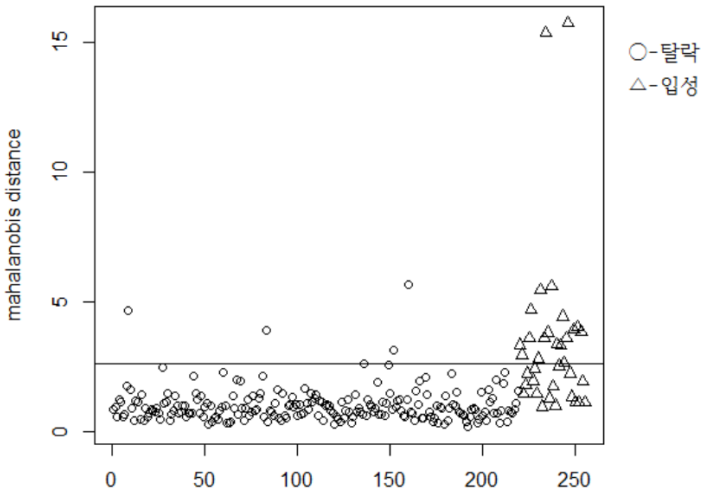

Figure 4.2. Scatterplot of MTS results using the Mahalanobis space based on the elimination group

비교의 역할을 하는 것에 비해 통계적 분류방법인 판별분석, 로지스틱 회귀분석, MTS방법은 명예의 전당의 입성, 탈락의 결과를 예측할 수 있다.

이 논문의 주된 목적은 통계적 분류방법에 MTS방법을 적용하여 그 결과를 소개하는 것이다. 이를 위해 기존의 분류방법 중 특히 로지스틱 회귀분석과 MTS방법의 분류결과를 비교하였으며 데이터의 속성상 어느 한 속성의 데이터가 뚜렷한 도형적 그룹으로 나타나지 않은 경우에는 더 효과적일 수 있음을 경험적으로 보여주고 있다.

또한 앞으로 명예의 전당의 입성 기준을 만족하는 새로운 선수가 있을 때 명예의 전당 지수보다 통계적 분류방법을 이용하여 판단을 한다면 좀 더 효과적인 판단 기준이 될 것이다. 그리고 이 논문에서는 타자의 공격부분에 대한 데이터만을 사용하였는데 수비부분에 대한 객관적인 지표들이 상대적으로 덜 발달하였고 각각의 수비 포지션도 평가 기준이 애매하기 때문에 제외했지만 앞으로 수비에 대한 객관적인 평가 기준을 개발하여 적용한다면 좀 더 정확한 분류를 할 수 있을 것으로 판단된다.

References

Hample, J. (2007). *Watching Baseball Smarter: A Professional Fan's Guide for Beginners, Semi-Experts, and Deeply Serious Geeks*, Vintage Books, USA.

Huh, M. and Yang, K. (2001). *Multivariate Data Analysis with SPSS*, Hannarae, Seoul.

James, B. (2001). *The New Bill James Historical Baseball Abstract*, Free Press, USA.

Johnson, R, A. and Wichern, D. W. (1982). *Applied Multivariate Statistical Analysis*, Prentice hall, USA.

Kwon, S. (2008a). *Multivariate Data Analysis and Applications*, Freedom Academy, Seoul.

Kwon, S. (2008b). *Regression analysis: Statistical Software SAS SPSS utilize the center*, Freedeom Academy, Seoul.

Lee, J. (2011). *Data Mining Using R, SAS and MS-SQL*, Freedom Academy, Seoul.

Marsh, N. (2007). *Baseball*, Murray Books, Australia.

Official website of Baseball Reference. (http://www.baseball-reference.com)

Park, C. (2012). A resetting scheme for process parameters using the Mahalanobis-Taguchi system, *The Korean Journal of Applied Statistics*, **25**, 589–603.

Taguchi, G., Chowdhury, S. and Wu, Y. (2005). *Taguchi's Quality Engineering Handbook*, Wiley, USA.

Taguchi, G. and Jugulum, R. (2000). New trends in multivariate diagnosis, *Indian Journal of Statistics, Series B*, **62**, 233–248.

Taguchi, G. and Jugulum, R. (2002). *The Mahalanobis-Taguchi Strategy*, Wiley, USA.

메이저리그 타자들의 명예의 전당 입성과 탈락에 대한 Mahalanobis-Taguchi System의 적용과 비교

김수환[a] · 박창순[a,1]

[a]중앙대학교 응용통계학과

(2012년 12월 31일 접수, 2013년 2월 6일 수정, 2013년 2월 7일 채택)

요 약

미국 프로야구(Major League Baseball) 명예의 전당의 입성과 탈락을 예측할 수 있는 여러 가지 통계적인 분류분석법을 실시하고 그 결과의 정확성을 비교하였다. 이를 위해 명예의 전당 가입 조건을 만족하는 타자들 중 1980년 이후 기록된 데이터의 17개의 독립변수를 사용하여 분류분석에서 널리 사용되는 기준으로 판별분석, 로지스틱 회귀분석과 상대적으로 최근에 제안된 Mahalanobis-Taguchi System(MTS)을 실시하여 비교하였다. 이 세 가지 방법 중 MTS가 상대적으로 더 나은 효율을 보였으며 이는 다변량 관측 값이 방향성이 없어 속성에 따른 도형적 그룹을 형성하지 못하는 경우에 효율적인 MTS의 특성에 의한 것으로 판단된다.

주요용0어: Mahalanobis-Taguchi System, 판별분석, 로지스틱 회귀분석, 분류정확도.

이 논문은 2011년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (156-756) 서울특별시 동작구 흑석동 221, 중앙대학교 응용통계학과, 교수.
E-mail: cspark@cau.cac.kr