응용통계연구 제 **21**권 *3*호, *2008*년, *pp. 455–468*

2007년 한국프로야구에서 도루성공모형

홍종선[1] 최정민[2]

요 약

야구경기의 승패에 영향을 미치는 중요한 요인으로 간주되는 도루의 성공모형을 개발하기 위하여 2007년 한국프로야구 기록자료를 바탕으로 로지스틱 회귀모형들을 제안한다. 또한 한국프로야구의 도루성공과 실패에 대해 판별분석을 실시하고 분류 기준값을 결정하였으며, 판별분석 분류표를 이용해 로지스틱 회귀분석과 판별분석의 효율성을 비교한다. 전체적인 모형의 정확도는 로지스틱 회귀모형이 판별분석보다 더 좋은 것으로 나타났고, 연속형 자료를 범주형으로 변환한 자료에 대한 로지스틱 회귀모형도 유사한 효율성을 갖고 있다.

주요용어: 로지스틱 회귀모형, 범주형, 분류 기준값, 분류표, 정확도, 판별분석.

1. 서론

1.1. 연구목적

한국프로야구는 출범 11년만인 2007년에 관중 수가 400만을 돌파했다. 전문가와 비전문가 즉, 남녀노소 누구나 쉽게 즐길 수 있는 스포츠로 변화한 한국프로야구는 2007년에 8개 구단별로 각각 126경기를 소화하면서 무수한 기록이 수립되었다. 수많은 경기기록 중 승패에 영향을 미치는 요인은 너무나 다양하지만 그 중에서도 중요한 요인으로 간주되는 도루에 대하여 연구한다. 현대 야구는 달리는 야구(running game)이기 때문에 팀의 훌륭한 주루 플레이는 상대팀을 당황케 하여 실수를 유발시키는데 보이지 않은 큰 공헌을 한다. 감각적인 주루 플레이로 팀의 분위기를 살리고, 반대로 아무 생각없는 주루 플레이가 그 팀의 분위기를 저하시키는 경우가 비일비재하다. 또한 주자가 1루에 있는 상황과 2루에 있는 상황은 큰 차이가 있다. 왜냐하면 주자가 1루에 있는 경우에 병살 플레이의 위험이 따르기 때문이다. 따라서 본 연구에서는 하나의 안타보다 더 값지고 경기의 흐름을 좌지우지할 수 있는 도루에 대한 통계적인 모형을 개발한다.

도루성공과 실패에 영향을 주는 요인으로는 일곱 장소의 경기장 요인, 홈경기 또는 원정경기의 영향, 주자의 아웃카운트에 대한 심리적 요인, 타자의 타격위치, 투수의 투구방향, 스트라이크 수, 볼 수, 2루심판 엄격도 등에 관한 요인과 더불어 투수, 포수, 타자, 주자 등에 관한 요인인 투수의 도루 허용률, 상대팀 수비율, 포수의 도루 저지율, 팀 총 도루수, 주자의 도루

1) (110-745) 교신저자. 서울시 종로구 명륜동 3-53, 성균관대학교 경제학부 통계학전공, 교수.
E-mail: cshong@skku.ac.kr
2) (110-745) 서울시 종로구 명륜동 3-53, 성균관대학교 응용통계연구소, 연구원.
E-mail: cjm6096@skku.edu

성공률 등을 고려하여 도루성공모형을 개발하고자 한다. 2007년 한국프로야구 총 경기의 기록 자료를 바탕으로 통계 모형을 이용하여 1루에 진출한 주자가 2루 도루를 성공하는가 또는 실패하는가를 예측할 수 있도록 모형 개발이 연구 목적이다.

한국프로야구 경기에 관한 연구는 한국의 통계학자들 사이에서도 활발하게 연구되어 많은 문헌들을 발견할 수 있다. 예를 들어 김용태와 이장택 (2005, 2006), 김응식 (2001), 김혁주 (2001, 2004, 2006), 오광모와 이장택 (2003), 조영석 등 (2007) 그리고 조영석과 조용주 (2003, 2004, 2005a, 2005b) 등이 있다. 체육학자들의 연구로는 김응식 (2002), 이장영과 강효민 (2001)의 투수의 경기력과 연봉에 대한 연구와 장건희 (1998), 손혁 (2004)의 투구와 구질에 관한 연구 그리고 타자의 타격과 투수의 투구 방향과의 관계를 중심으로 박순욱 (2006) 등이 야구의 투수에 대한 연구를 하였으나 야구경기에서 주자들의 도루성공에 대한 연구는 미비한 실정이다.

분석하기 위해 수집된 자료 설명을 한 뒤, 2절에서는 일변량 자료분석을 한다. 설명변수들을 범주형 변수와 연속형 변수로 나누어 탐색적 분석을 한다. 3절에서는 2007년 한국프로야구 도루 성공을 가장 잘 분석하고 예측하는 로지스틱 회귀모형을 제안한다. 제안한 모형의 설명변수 중 수집하기 어려운 연속형 변수를 범주형으로 변환시켜 로지스틱 회귀분석을 실시한 결과를 4절에서 비교 토론한다. 5절에서는 제안한 로지스틱 회귀모형과 판별분석으로 분류한 결과를 비교 분석하고, 6절에서 결론을 맺는다.

1.2. 자료설명

본 연구는 2007년 한국프로야구의 도루성공모형에 관한 연구를 하기 위하여 2007년 4월 6일부터 2007년 10월 19일까지 한국프로야구의 각 8팀별 126경기를 종합한 총 504경기의 기록 자료를 바탕으로 도루의 성공 여부에 관련이 있다고 간주되는 표 1.1과 같은 변수의 자료(총 1,251건)를 분석하고자 한다. 주자, 포수, 투수등에 관한 자료는 가능한 많은 정보를 활용하기 위해 이전 연도(2005, 2006년)에 대한 자료도 투입하였다. 나머지 경기상황에 대한 자료는 모두 2007년 자료이다.

수집한 자료에서 범주형 변수 중 범주 수가 많은 볼 카운트(ball count)에 대하여는 0스트라이크 0볼부터 2스트라이크 3볼까지의 변수값을 그대로 사용하여 볼 카운트의 12가지 범주를 그대로 모형에 투입한 방법과 12가지 볼 카운트 중 볼이 더 많은 경우와 그렇지 않은 경우로 구분한 방법과 마지막으로 볼 카운트 변수를 스트라이크와 볼이라는 두 변수로 분리시키는 방법을 고려해 보았는데 이 중에서 볼 카운트를 스트라이크와 볼이라는 두 개의 변수로 분리시켜 모형에 투입하는 방법이 모형의 유의성이 가장 큰 것을 발견하였다. 따라서 볼 카운트는 '스트라이크 수'($X6$), '볼 수'($X7$)의 두 개로 분리시켜 모형에 적용하였다. 그리고 2루 도루의 성공 여부는 2루 심판에 따라 크게 차이가 난다는 여론을 반영하여 2루 심판에 대한 변수를 고려하였다. 2루 심판의 범주 수는 24명이었고 각 심판의 도루 성공, 실패의 비율을 고려해 심판의 아웃시킬 확률을 산출하였다. 수집된 자료에서 2루 심판 24명 모두를 범주 수준으로 간주하여 모형에 투입한 방법보다는 2루 심판의 각각 도루 아웃시킬 확률을 산출해서 0.2미만의 심판을 A, 0.2이상 0.3미만은 B, 0.4이상은 C로 범주화시켜 2루 심판의 엄격한 정도를 3가지 범주로 변환하여 '2루 심판 엄격도' 변수($X8$)를 생성하여 모형에 적용하였다.

표 1.1: 도루성공모형의 자료

변수		변수값
반응변수	도루성공/실패(Y)	1: 도루성공(821건), 2: 도루실패(430건)
설명변수(범주형)	경기장($X1$)	1: 잠실구장, 2: 인천구장, 3: 수원구장, 4: 대전구장, 5: 광주구장, 6: 대구구장, 7: 부산구장
	주자의 홈/ 어웨이($X2$)	1: 홈경기, 2: 원정경기
	아웃카운트($X3$)	0, 1, 2: 아웃카운트
	타자의 타격위치($X4$)	1: 우타자, 2: 좌타자, 3: 스위치히터
	투수의 투구방향($X5$)	1: 우완투수, 2: 좌완투수, 3: 언더핸드(우완)
	스트라이크 수($X6$)	0, 1, 2
	볼 수($X7$)	0, 1, 2, 3
	2루 심판 엄격도($X8$)	A, B, C (2007년 자료)
	팀 총 도루 수($X9$)	2005, 2006, 2007 팀별 총 도루 수(8범주)
	상대팀 수비율($X10$)	2006, 2007 팀별 수비율(8범주)
설명변수(연속형)	포수의 도루 저지율($X11$)	2006, 2007 포수의 도루 저지율
	투수의 도루 허용률($X12$)	2006, 2007 투수의 도루 허용률
	주자의 도루 성공률($X13$)	2006, 2007 주자의 도루 성공률

*자료 참고 및 출처: 스포츠투아이 주식회사(www.sports2i.com)
통계로 즐기는 프로야구(www.istat.co.kr)

2. 일변량 자료분석

2.1. 범주형 자료분석

반응변수와 각 설명변수의 관계를 파악하기 위하여 탐색적 자료분석(exploratory data analysis)을 하였다. 범주형 설명변수에 대하여는 막대그래프를 작성하고 연속형 설명변수에 대하여는 산점도를 이용해 자료를 탐색적으로 분석하고, 반응변수와 각 설명변수와의 유의성을 알아보기 위하여 범주형 설명변수는 교차분석을, 연속형 설명변수는 단순로지스틱 회귀분석을 실시한다.

범주형 설명변수와 도루 성공률인 반응변수 간의 원 막대그래프를 작성한 그림 2.1을 살펴보자. 수평축은 각 변수와 범주 수준을 나타내고 수직축은 실제 기록된 도루성공비율(빗금친 부분)과 실패비율을 구현하였다. 각 변수의 범례는 표 1.1을 참고할 수 있다. 표 2.1에 반응변수와의 교차분석 결과인 카이제곱 통계량값과 대응하는 p-값을 나타내었다. 범주형 설명변수와 반응변수 간의 유의성을 분석해 본 결과, 투수의 투구방향($X5$)과 볼 수($X7$), 2루 심판의 엄격도($X8$), 팀 총 도루수($X9$), 상대팀 수비율($X10$)등의 변수들이 도루성공/실패(Y)에 유의한 영향을 미치는 설명변수라는 것을 알 수 있다.

2.2. 연속형 자료분석

다음 그림 2.2는 연속형 설명변수와 반응변수 간의 산점도이다. 수평축은 각 변수를 나타내고 수직축은 실제 기록된 도루 성공률을 나타내었다.

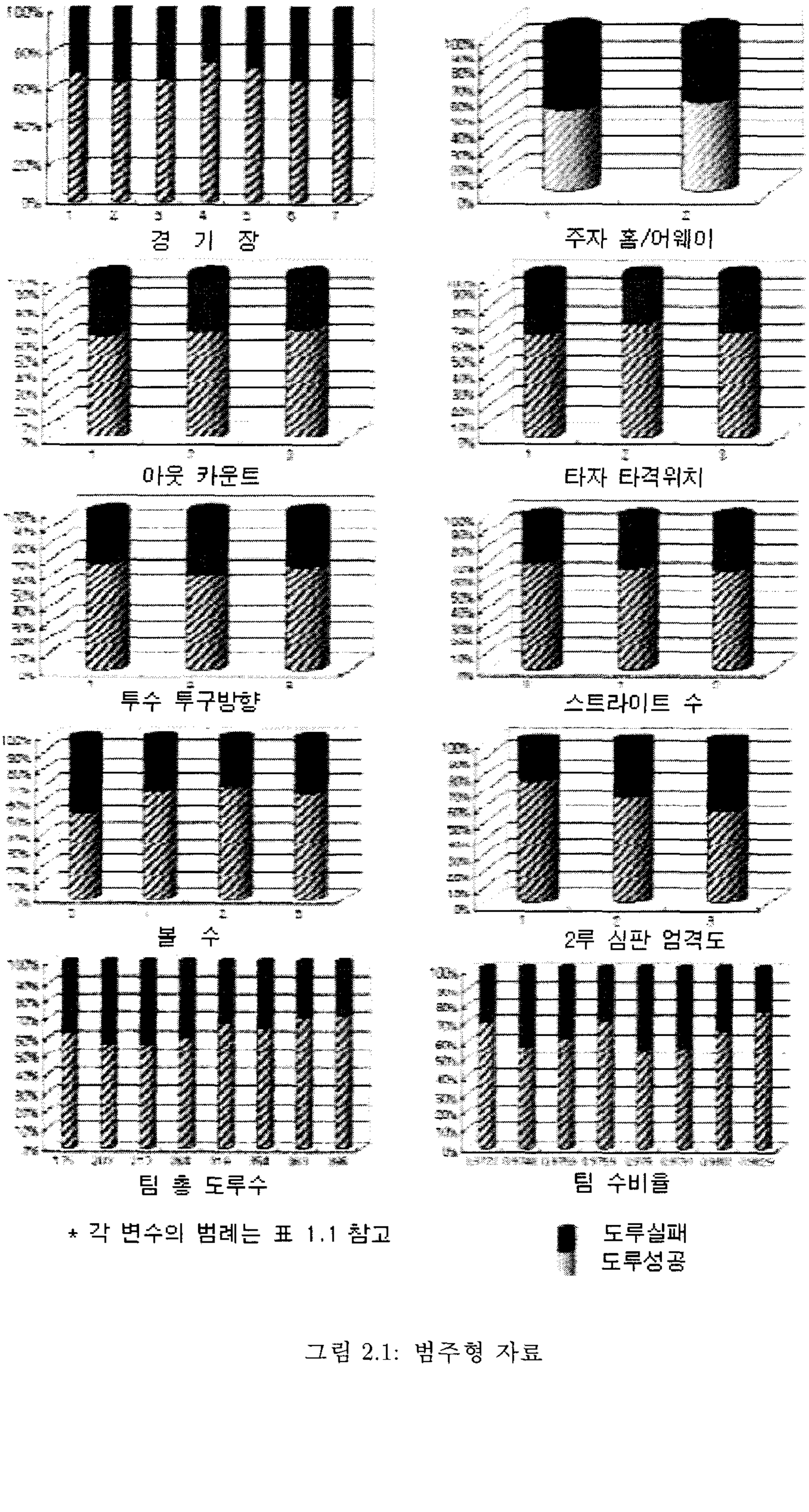
경 기 장
주자 홈/어웨이
아웃 카운트
타자 타격위치
투수 투구방향
스트라이트 수
볼 수
2루 심판 엄격도
팀 총 도루수
팀 수비율
* 각 변수의 범례는 표 1.1 참고
도루실패
도루성공

그림 2.1: 범주형 자료

표 2.1: 반응변수와의 카이제곱 검정 결과

변수목록	χ^2	p-값
경기장($X1$)	11.3271	.078
주자 홈/어웨이($X2$)	2.6653	.102
아웃 카운트($X3$)	2.0288	.362
타자 타격위치($X4$)	5.4670	.065
투수 투구방향($X5$)	6.0517	.048
스트라이크 수($X6$)	2.5233	.283
볼 수($X7$)	14.3000	.002
2루심판 엄격도($X8$)	21.3388	.001
팀 총 도루수($X9$)	15.4764	.030
상대팀 수비율($X10$)	27.1250	.001

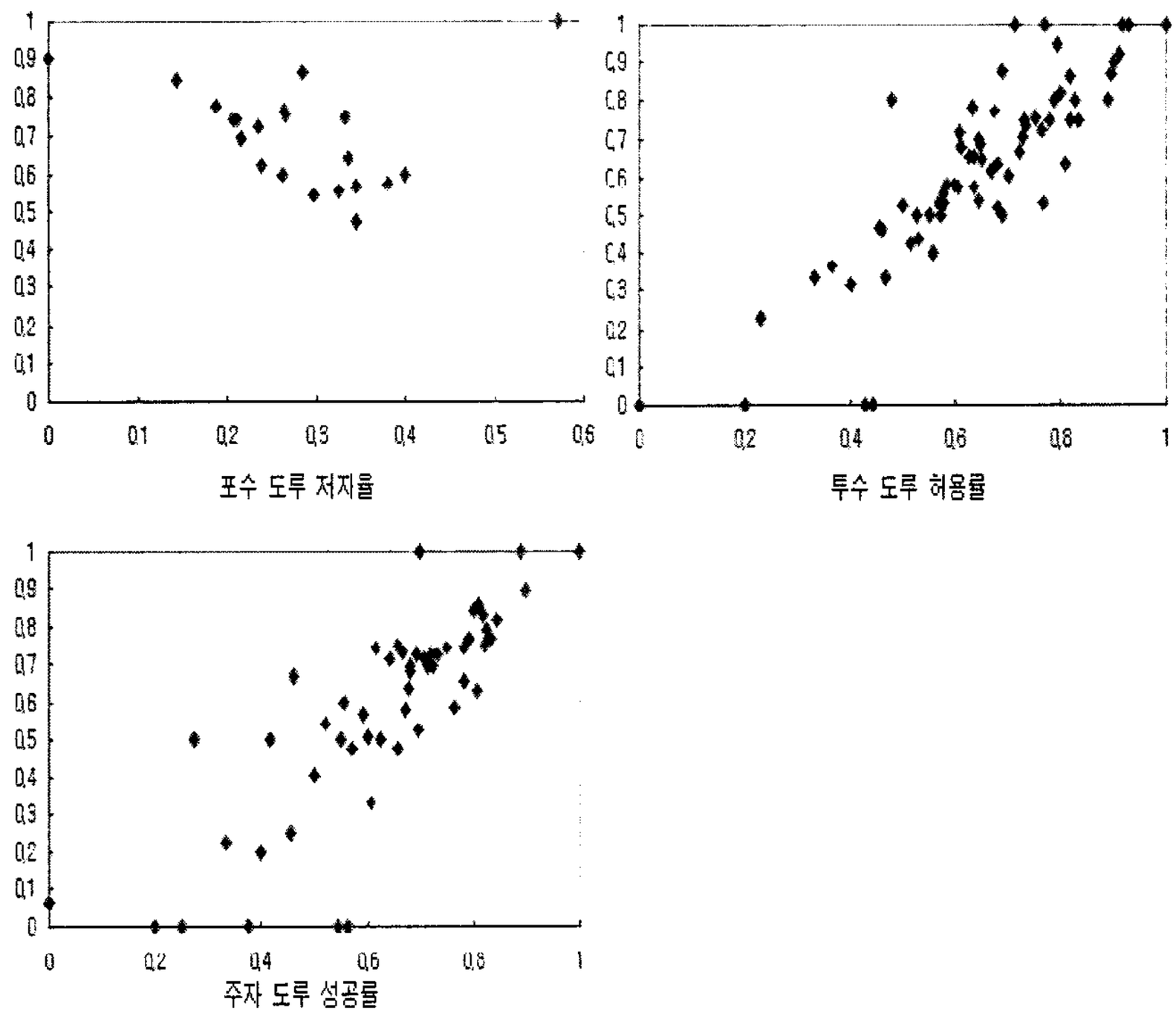

그림 2.2: 연속형 자료

연속형 설명변수와 반응변수의 단순 로지스틱 회귀모형의 유의성을 정리한 표 2.2를 통해서, 포수의 도루 저지율($X11$), 투수의 도루 허용률($X12$), 주자의 도루 성공률($X13$)의 회귀모형이 반응변수(Y)에 유의한 영향을 미치는 연속형 변수임을 알 수 있다.

표 2.2: 반응변수와의 로지스틱 회귀모형 검정 결과

변수목록	χ^2	p-값
2006, 2007 포수의 도루 저지율($X11$)	26.0383	.001
2006, 2007 투수의 도루 허용률($X12$)	165.9408	.001
2006, 2007 주자의 도루 성공률($X13$)	127.7040	.001

표 3.1: 단계선택방법 결과

단계	투입	제거	자유도	χ^2	p-값
단계1	투수의 도루 허용률($X12$)	.	1	155.8560	.001
단계2	주자의 도루 성공률($X13$)	.	1	114.6472	.001
단계3	2루 심판 엄격도($X8$)	.	2	16.7779	.001
단계4	포수의 도루 저지율($X11$)	.	1	6.8023	.009
단계5	상대팀 수비율($X10$)	.	7	15.1565	.034
단계6	볼 수($X7$)	.	3	9.4808	.023

표 3.2: 최종 로지스틱 회귀모형

회귀계수		추정값	표준오차	χ^2(Wald)	p-값	오즈비
상수항		−44.7362	0.7164	43.7024	0.001	
볼 수($X7$)	0	−40.4442	0.1509	8.6634	0.003	0.532
	1	0.0518	0.1114	0.2157	0.642	0.873
	2	0.2049	0.1166	3.0868	0.078	1.018
2루 심판 엄격도($X8$)	1	0.4800	0.1291	13.8187	0.001	2.247
	2	−0.1501	0.0958	2.4574	0.117	1.197
상대팀 수비율($X10$)	현대(1)	−0.7431	0.2207	11.3382	0.001	0.567
	삼성(2)	0.5430	0.2331	5.4257	0.019	2.052
	한화(3)	−0.2001	0.1519	1.7366	0.187	0.976
	두산(4)	0.1103	0.2056	0.2880	0.591	1.331
	SK(5)	−0.1171	0.1923	0.3710	0.542	1.061
	LG(6)	0.2054	0.2272	0.8174	0.365	1.464
	KIA(7)	0.3775	0.2090	3.2630	0.070	1.739
포수 도루 저지율($X11$)		−6.7282	1.7450	14.8580	0.001	0.001
투수 도루 허용률($X12$)		5.6368	0.5498	105.1024	0.001	280.564
주자 도루 성공률($X13$)		5.4165	0.5660	91.5771	0.001	225.091

가능도비 카이제곱 통계량 : 333.43*** (자유도: 15)

***: 유의확률 <.001

3. 로지스틱 회귀모형

표 1.1에 나열한 15개의 변수를 고려한 로지스틱 회귀모형으로부터 단계선택방법(SLE: 0.05, SLS: 0.05)를 이용해 변수를 선택하여 최량의 로지스틱 회귀모형을 선정한 결과를 표 3.1에 나타내었다.

최종모형으로 선택된 변수들은 투수의 도루 허용률($X12$), 주자의 도루 성공률($X13$), 2루

표 3.3: 로지스틱 회귀분석 결과

단계	분류기준값	TP	TN	FP	FN	민감도	1-특이도	TP+TN
6	0.578	676	253	177	145	0.82439	0.41162	929
6	0.589	668	260	170	153	0.81463	0.39534	929
6	0.584	671	257	173	150	0.81829	0.40232	929
6	0.582	672	256	174	149	0.81951	0.40465	928
6	0.443	749	179	251	72	0.91341	0.58372	928
6	0.441	750	178	252	71	0.91463	0.58604	928
6	0.590	667	260	170	154	0.81341	0.39534	927
6	0.585	670	257	173	151	0.81707	0.40232	927
6	0.584	671	256	174	150	0.81829	0.40465	927
6	0.580	672	255	174	150	0.81951	0.40697	927

표 3.4: 로지스틱 회귀분석에 의한 분류표

		예측		합
		도루성공	도루실패	
실제	도루성공	676 (82.4%)	145 (17.6%)	821 (100.0%)
	도루실패	177 (41.1%)	253 (58.9%)	430 (100.0%)
합		853 (68.2%)	398 (31.8%)	1251 (100.0%)

심판 엄격도($X8$), 포수의 도루 저지율($X11$), 상대팀 수비율($X10$), 볼 수($X7$) 등 총 6개의 변수이며, 최종 로지스틱 회귀모형은 표 3.2와 같다.

가능도비 카이제곱 검정(likelihood ratio chi square test)결과 검정통계량 값은 333.43이고 p-값은 0.001로 0에 가까운 매우 작은 값을 갖기 때문에 모형의 유의성이 타당하다고 판단된다. 오즈비(odds ratio)는 어떤 사건이 발생되지 않을 확률에 대한 발생될 확률의 비율을 의미하는데, 본 연구에서는 도루실패 확률에 대한 도루성공 확률의 비율을 의미한다고 볼 수 있다. 표 3.2에서 보면 볼 수($X7$)가 증가할수록 오즈비는 높아진다. 따라서 볼 수($X7$)가 증가할수록 도루실패에 대한 도루성공의 비율은 높아진다고 할 수 있다. 이와는 반대로 2루심판의 엄격도($X8$)가 증가할수록 오즈비는 감소하므로 도루실패에 대한 도루성공의 비율은 낮아진다는 것을 파악할 수 있다. 다음은 단계적 변수선택 방법을 이용하여 선정된 최량의 로지스틱 회귀모형의 결과 중에서 최종 단계인 여섯번째 단계에서의 로지스틱 회귀모형 분류표를 출력한 결과를 표 3.3에 나타내었다.

분류 기준값에 따라 분류 확률이 각각 다르기 때문에 최상의 분류 기준값을 찾기 위해 연구하였다. 표 3.3으로부터 가장 좋은 분류 정확도를 갖는 분류 기준값은 0.578로 TP(True Positive: 정확히 도루성공을 예측한 경우)는 676개, TN(True Negative: 정확히 도루실패를 예측한 경우)는 253개로 분류되는 것을 알 수 있다. TP+TN값이 높을수록 정확하게 분류한 경우가 많다는 것을 나타낸다. 즉 전체 경우 중 1251개 중 929개를 정확히 분류하여 약

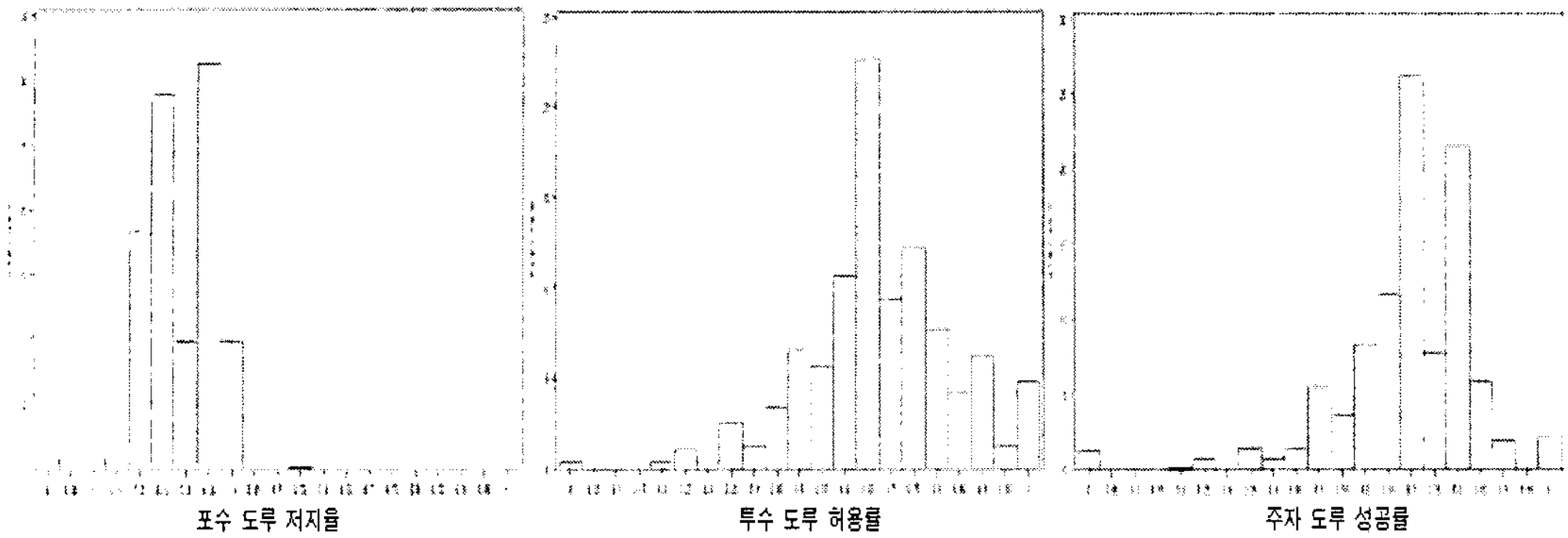

그림 4.1: 포수 도루 저지율, 투수 도루 허용률, 주자 도루 성공률의 분포

74.2%의 정확도를 나타낸다. 이 분석의 결과를 바탕으로 실제와 예측한 도루성공여부에 관한 분류표로 나타내면 표3.4와 같다. 표 3.4에 대하여는 4절의 표 4.3과 같이 5절의 표 5.2를 설명하면서 비교 토론한다.

4. 범주형 변수에 대한 로지스틱 회귀모형

3절에서 논의한 최종 로지스틱 회귀모형의 연속형 설명변수 중 '포수 도루 저지율'($X11$), '투수 도루 허용률'($X12$), '주자 도루 성공률'($X13$) 변수의 특성 중의 하나는 세 변수 모두 양수 값의 첨도(1.91, 1.12, 5.67)와 음수 값의 왜도(-0.61, -0.36, -1.66)를 가진 분포를 따르고 그림 4.1에서와 같이 세 변수 각각의 중위수(0.29, 0.67, 0.79) 주변에 많은 값들이 몰려있으며 약간 왼쪽으로 치우쳐 있다.

그림 4.1과 같은 형태로 균등한 분포를 보이지 않고 치우침이 있다고 판단되어 비전문가들이 이해하기 쉽도록 그리고 정확한 값을 모른다 하더라도 소수의 범주수준으로는 변환 가능하기 때문에 연속형 변수를 범주형 변수로 변환하여 로지스틱 회귀분석을 실시하였다. 설명변수의 값의 범위인 0과 1 사이를 등간격의 세 구간부터 다섯 구간으로 범주화하여 로지스틱 회귀분석한 결과와 0과 1 사이를 세 구간부터 다섯 구간의 등비율로 범주화하여 로지스틱 회귀분석한 결과를 비교해본 결과, 0.25씩 동일한 비율로 분할하여 네 수준의 범주화한 자료에 대한 로지스틱 회귀모형이 제일 좋은 결과를 나타내고 있으며, 단계적 변수선택 방법을 사용하여 얻은 최량 로지스틱 회귀모형은 3절에서 설정한 최종 모형의 설명변수들 중에서 상대팀 수비율($X10$)변수가 제외된 다섯 개의 설명변수로 구성되었으며 이를 표 4.1에 정리하였다.

가능도비 카이제곱 통계량의 값에 대응하는 p-값은 0.001보다 작은 값을 갖기 때문에 3절에서의 모형과 동일하게 유의하다고 판단된다. 모든 설명변수가 범주형 변수인 변환된 자료에 대하여 단계적 변수선택 방법을 사용하여 구한 최량의 로지스틱 회귀모형의 분석결과를 표 4.2에 나타내었다.

표 4.2로부터 가장 좋은 분류 정확도를 갖는 분류 기준값은 0.551로 TP는 681개, TN는 244개로 분류되는 것을 알 수 있다. 즉 전체 경우 1251개 중 925(TP+TN)개를 정확히 분류

표 4.1: 범주형 자료의 최종 로지스틱 회귀모형

회귀계수		추정값	표준오차	$\chi^2(Wald)$	p-값	오즈비
상수항		0.8370	0.088	90.1513	0.001	
볼 수($X7$)	0	−0.4109	0.1458	7.9490	0.004	0.560
	1	0.0668	0.1082	0.3812	0.536	0.902
	2	0.1746	0.1135	2.3659	0.124	1.005
2루 심판 엄격도($X8$)	1	0.5593	0.1255	19.8735	0.001	2.697
	2	−0.1267	0.0933	1.8430	0.174	1.358
포수 도루 저지율($X11$)	1	0.0783	0.1150	0.4645	0.495	1.357
	2	0.3341	0.1409	5.6236	0.017	1.753
	3	−0.1854	0.1090	2.8964	0.088	1.042
투수 도루 허용률($X12$)	1	−0.1854	0.1143	63.6711	0.001	0.117
	2	−0.9124	0.1132	3.4233	0.064	0.237
	3	−0.1091	0.1203	0.8229	0.364	0.262
주자 도루 성공률($X13$)	1	−1.0033	0.1138	77.7417	0.001	0.184
	2	0.0304	0.1250	0.0591	0.807	0.517
	3	0.2826	0.1107	6.5164	0.010	0.665
가능도비 카이제곱 통계량 : 276.49*** (자유도:14)						

***: 유의확률 <.001

표 4.2: 로지스틱 회귀분석 결과

단계	분류기준값	TP	TN	FP	FN	민감도	1-특이도	TP+TN
5	0.551	681	244	186	140	0.83048	0.43255	925
5	0.553	680	244	186	141	0.82926	0.43255	924
5	0.542	684	240	190	137	0.83414	0.44186	924
5	0.546	682	241	189	139	0.83170	0.43953	923
5	0.543	683	240	190	138	0.83292	0.44186	923
5	0.541	685	238	192	136	0.83536	0.44651	923
5	0.556	678	244	186	143	0.82682	0.43258	922
5	0.541	685	236	194	136	0.83536	0.45116	921
5	0.526	692	229	201	129	0.84390	0.46744	921
5	0.541	685	235	195	136	0.83536	0.45348	920

하여 약 73.9%의 정확도를 나타내며 표 4.2의 분류표로부터 확인된다. 3절에서 얻은 최량 로지스틱 회귀모형으로 구한 분류 정확도 74.2%와 동등한 정확도를 보여준다. 범주형으로 변환한 자료에 대한 분석결과를 바탕으로 얻은 분류표는 표 4.3과 같다.

5. 판별분석과 로지스틱 회귀모형과의 비교

로지스틱 회귀분석과 비슷한 개념의 판별분석을 실시하여 앞에서 연구한 로지스틱 회귀모형과 비교하고자 한다. 판별분석의 결과는 표 5.1에 나타내었다. 표 5.1을 정리하여 첫번째 판별함수인 도루성공에 대한 판별함수와 두번째 판별함수인 도루실패에 대한 판별함수는 각

표 4.3: 범주형 자료의 로지스틱 회귀분석에 의한 분류표

		예측		합
		도루성공	도루실패	
실제	도루성공	681 (82.9%)	140 (17.1%)	821 (100.0%)
	도루실패	186 (43.3%)	244 (56.7%)	430 (100.0%)
합		867 (69.3%)	384 (30.7%)	1251 (100.0%)

표 5.1: 판별함수 결과

변수목록	Target(1)	Target(2)
상수항	−64.254	−58.281
경기장($X1$)	0.925	0.941
홈/어웨이($X2$)	5.910	5.753
아웃카운트($X3$)	2.573	2.389
타자유형($X4$)	4.991	4.936
투수유형($X5$)	3.141	3.137
스트라이크($X6$)	1.449	1.546
볼($X7$)	1.500	1.344
2루 심판 엄격도($X8$)	6.236	6.638
팀 총 도루수($X9$)	0.844	0.845
상대팀 수비율($X10$)	0.427	0.393
포수 도루 저지율($X11$)	84.025	86.982
투수 도루 허용률($X12$)	40.719	35.592
주자 도루 성공률($X13$)	36.266	31.052

각 다음과 같고, 전체에 대한 분류표(classification table)는 표 5.2에 나타내었다.

$$\begin{aligned}\text{Target}(1) = &-64.254 + 0.925 \times X_1 + 5.910 \times X_2 + 2.573 \times X_3 + 4.991 \times X_4 \\ &+ 3.141 \times X_5 + 1.449 \times X_6 + 1.500 \times X_7 + 6.236 \times X_8 \\ &+ 0.844 \times X_9 + 0.427 \times X_{10} + 84.025 \times X_{11} + 40.719 \times X_{12} \\ &+ 36.266 \times X_{13} \\ \text{Target}(2) = &-58.281 + 0.941 \times X_1 + 5.753 \times X_2 + 2.389 \times X_3 + 4.936 \times X_4 \\ &+ 3.137 \times X_5 + 1.546 \times X_6 + 1.344 \times X_7 + 6.638 \times X_8 \\ &+ 0.845 \times X_9 + 0.393 \times X_{10} + 86.982 \times X_{11} + 35.592 \times X_{12} \\ &+ 31.052 \times X_{13}\end{aligned}$$

표 5.2는 판별분석으로 정리된 분류표이다. 도루성공을 정확히 예측한 확률은 74.7% 이고 도루실패를 정확히 예측한 확률은 66.1%로 나타났다. 전체에 대한 분류 정확도는 $(614 + 284)/1251 = 0.7178$로 약 71.7%를 정확하게 분류한 것을 알 수 있다. 3절에서 논의한 로지스

표 5.2: 판별분석에 의한 분류표

		예측		합
		도루성공	도루실패	
실제	도루성공	614 (74.7%)	207 (25.3%)	821 (100.0%)
	도루실패	146 (33.9%)	284 (66.1%)	430 (100.0%)
합		760 (60.8%)	491 (39.2%)	1251 (100.0%)

틱 회귀모형의 분류 정확도는 74.2%이며, 4절에서 논의한 범주형 자료에 대한 로지스틱 회귀모형의 분류 정확도는 조금 낮으나 동등한 정확도를 보여주는 73.9%와 비교하여 판별분석의 분류 정확도는 71.7%로 로지스틱 모형의 정확도보다 차이가 발생한다. 그러므로 로지스틱 회귀모형과 범주형 자료에 대한 로지스틱 회귀모형의 분류 정확도는 차이가 없고 판별분석의 분류 정확도보다 높다고 판단할 수 있다. 참고로 3절에서 논의한 유의한 변수만으로 판별분석을 실시해 보았으나 표 5.2보다 정확도 측면에서 조금 낮은 결과가 발생하였기 때문에 본문에 추가하지 않았다.

6. 향후 연구과제 및 결론

본 연구에서는 2007년 한국프로야구의 도루성공과 실패 모형을 분석과 예측을 위해 로지스틱 회귀모형을 이용하여 자료를 설명하고 분석하였다. 야구자료의 특성상 도루에 영향을 미치는 요인이 많이 존재하지만, 구단의 개인적인 문제와 스포츠 종목의 특성상 완벽하고 정확한 자료를 수집하는 것에 어려움이 많았다. 또한 도루기록 자료에서도 도루성공 횟수(821건)가 도루실패 횟수(430건)에 비해 두배정도 많기 때문에 로지스틱 회귀모형의 선형성 문제로 인한 도루실패에 대한 분류 정확도가 다소 떨어지는 문제가 발생하기도 하였다. 특히 본 논문에서 논의한 연속형 설명변수들(포수 도루 저지율, 투수 도루 허용률, 주자 도루 성공률)은 구단보다는 개인적인 질적 자료로 공식적으로 최근의 자료를 수집하기 쉽지 않은 변수들이다. 또한 각각의 설명변수들을 네 등분한 범위로 나누어 범주형 변수로 변환시켜 분석한 결과는 원자료를 분석한 결과와 매우 유사하였다. 특히 포수 도루 저지율($X11$), 투수 도루 허용률($X12$), 주자 도루 성공률($X13$)은 일반 관중들이 수집하기 어려운 자료이다. 이 변수들의 값을 정확히 모르더라도 네 등급(예를 들어 상, 중상, 중하, 하)으로 범주화한다면 관중들도 어렵지 않게 측정할 수 있기 때문에 분석하는 것이 용이해진다. 따라서 4절에서 논의한 범주형 자료를 이용한 로지스틱 회귀모형을 이용하면, 일반 관중들도 감독과 선수와 같은 높은 수준의 경기 분석을 하면서 경기를 즐길 수 있겠다.

도루실패 건수가 도루성공 건수에 비해 현저하기 작기 때문에, 성공의 분류 정확도는 82.4%나 실패의 분류 정확도는 58.9%로 낮은 값을 나타낸다(표 3.4 참조). 이를 위해 판별분석을 실시한 결과, 성공의 분류 정확도는 74.7%로 낮아지나 실패의 분류 정확도는 66.1%로 높은 값을 나타낸다(표 5.2 참조). 따라서 선형성 문제로 인한 도루실패 예측 정확도를 어느

정도 보완하였으나, 전체적으로 로지스틱 회귀분석의 분류 정확도(74.2%)보다 판별분석의 분류 정확도(71.7%)가 낮은 값을 보여준다. 그러므로 야구의 도루성공모형으로는 로지스틱 회귀모형이 자료를 잘 설명한다는 것을 적용해 보았으며, 특히 2007년 한국프로야구의 도루성공과 실패의 예측을 위하여 로지스틱 회귀모형을 제안한다. 또한 수집하기 어려운 연속형 설명변수 자료는 네 등급으로 수준으로 나누어 변환시킨 범주형 자료를 사용한 로지스틱 회귀모형도 자료를 설명하는데 충분하다는 것을 보일 수 있었다. 본 연구에서 제안한 모형을 잘 활용하면, 일반 관중들도 감독과 선수 수준에 못지않은 야구 경기 분석을 할 수 있어 경기를 관람하는데 많은 정보를 제공할 수 있겠다.

참고문헌

김용태, 이장택 (2005). 한국프로야구에 적당한 득점 추정측도에 관한 연구, <한국자료분석학회지>, **7**, 2289–2302.

김용태, 이장택 (2006). 한국프로야구에서의 승률추정에 관한 연구, <한국자료분석학회지>, **8**, 857–869.

김응식 (2001). 한국프로야구 선수의 경기력과 연봉과의 관계, <한국스포츠사회학회지>, **14**, 15–24.

김응식 (2002). 한국프로야구 투수의 경기력과 연봉과의 관계, <한국스포츠사회학회지>, **15**, 95–104.

김혁주 (2001). 스포츠에서 부적절하게 사용되고 있는 통계적 개념에 관한 소고, <응용통계연구>, **14**, 201–210.

김혁주 (2004). 한국프로야구와 프로축구의 순위 결정 기준은 통계적 관점에서 합리적인가?, <한국자료분석학회지>, **6**, 1767–1775.

김혁주 (2006). 스포츠에서 쓰이는 통계적 개념의 합리성에 관하여: 야구와 축구를 중심으로, <한국체육측정평가학회지>, **8**, 53–66.

박순욱 (2006). 야구경기 상황별 투구내용에 따른 결과 분석, <명지대학교 스포츠기록분석전공, 석사학위 논문>.

손혁 (2004). 프로야구 투수 유형과 구질과의 관계, <고려대학교 체육교육전공, 석사학위 논문>.

오광모, 이장택 (2003). 데이터마이닝을 이용한 한국프로야구 선수들이 연봉에 관한 모형연구, <한국스포츠사회학회지>, **16**, 295–309.

이장영, 강효민 (2001). 한국프로야구 투수의 경기수행과 연봉책정의 관계, <한국스포츠사회학회지>, **14**, 115–125

장건희 (1998). 프로야구 투수 유형별 투구내용 분석, <건국대학교 사회교육전공, 석사학위논문>.

조영석, 조용주 (2003). 한국프로야구에서 Beane Count 적용에 관한 연구, <한국자료분석학회지>, **5**, 649–658.

조영석, 조용주 (2004). 2003시즌 한국프로야구에서 WHIP가 방어율에 미치는 영향에 관한 연구, <한국자료분석학회지>, **6**, 1415–1424.

조영석, 조용주 (2005a). 한국프로야구에서 OPS와 득점에 관한 연구, <한국자료분석학회지>, **7**, 221–231.

조영석, 조용주 (2005b). 한국프로야구에서 득점과 실점을 이용한 승률 추정에 관한 연구, <한국자료분석학회지>, **7**, 2303–2312.

조영석, 조용주, 신상근 (2007). 한국프로야구에서 승패 추정에 관한 연구, <한국자료분석학회지>, **9**, 501–510.

[2008년 3월 접수, 2008년 5월 채택]

Steal Success Model for 2007 Korean Professional Baseball Games

Chong Sun Hong[1)] Jeong Min Choi[2)]

ABSTRACT

Based on the huge baseball game records, the steal plays an important role to affect the result of games. For the research about success or failure of the steal in baseball games, logistic regression models are developed based on 2007 Korean professional baseball games. The analyses of logistic regression models are compared of those of the discriminant models. It is found that the performance of the logistic regression analysis is more efficient than that of the discriminant analysis. Also, we consider an alternative logistic regression model based on categorical data which are transformed from uneasy obtainable continuous data.

Keywords: Category, logistic regression model, discriminant analysis, classification criterion, classification table.

1) Corresponding author. Professor, Dept. of Statistics, Sungkyunkwan University, 3-53, Myungryun-dong 3, Jongro-gu, Seoul 110-745, Korea.
E-mail: cshong@skku.ac.kr
2) Researcher, Research Institute of Applied Statistics, Sungkyunkwan University, 3-53, Myungryun-dong 3, Jongro-gu, Seoul 110-745, Korea.
E-mail: cjm6096@skku.edu