Journal of the Korean Data & Information Science Society
2017, **28**(2), 317–326

http://dx.doi.org/10.7465/jkdi.2017.28.2.317
한국데이터정보과학회지

한국프로야구에서 투수 연봉에 영향을 주는 요인

이장택[1]

[1]단국대학교 응용통계학과

접수 2017년 2월 11일, 수정 2017년 3월 19일, 게재확정 2017년 3월 27일

요 약

한국프로야구에서 투수들의 연봉을 결정하는 중요한 요인들을 선형회귀모형을 통해서 살펴본다. 제안된 모형을 이용하여 투수들의 경기력과 연봉간의 패턴을 분석할 수 있으며, 구단 별로 차이점도 알 수 있다. 사용된 데이터는 2010년부터 2015년까지의 투수 기록과 다음 해의 연봉 자료를 이용하였으며, 고려된 설명변수들은 해당연도, 팀의 종류, 게임 수, 평균자책점, 수비무관 평균자책점, 이닝당 안타 및 볼넷 허용률, 대체선수 대비 승리기여도, 선발출장 게임의 수, 승, 패, 세이브, 투구 이닝 수, 자유계약선수 여부, 나이, 경험연수이며 반응변수로는 연봉에 로그를 취한 로그연봉을 사용하였다. 그 결과 선발투수이며 경기수가 많고 승수가 많은 투수들에게 많은 연봉이 지급되고 있고 투수의 고유능력을 평가할 수 있는 기록들은 반영이 작게 되고 있음을 확인할 수 있었으며 연구의 결과는 연봉 결정에 중요한 참고자료로 활용될 수 있을 것으로 간주된다.

주요용어: 경기력, 선발투수, 투수 연봉, 한국프로야구, 회귀모형.

1. 서론

프로야구에서 투수가 차지하는 비중은 날이 갈수록 높아지고 있는 것이 오늘날의 추세이다. 확실한 투수가 없는 경우에 한국시리즈에서 우승할 확률은 거의 없으며 투수들은 장기 페넌트레이스를 운영하는 데 있어서 전력의 핵심이다. 따라서 매우 중요한 위치에 있는 투수들의 연봉은 반드시 능력을 제대로 반영하여 보다 객관화해서 정해야만 한다. 프로야구 구단에서 연봉 책정 시에 가장 많은 비중을 차지하는 부분은 구단의 고과 평점으로 알려져 있다 (Seung과 Kang, 2012). 각 구단마다 약간의 차이는 있겠지만 선수가 팀에서 차지하는 비중, 기여도, 인기 및 경력 등을 종합적으로 고려된 고과 평점을 바탕으로 연봉을 정한다고 한다. 그러나 한국프로야구 (Korean Baseball Organization; KBO)에도 고과 평점으로 객관적인 판단을 한다지만 연봉이 많아질수록 더욱 연봉 협상과 관련된 문제가 많이 발생하며 미국이나 일본에 비해 구단의 경제적 형편도 썩 좋은 편이 아니기 때문에 구단과 선수들 사이에 서로가 수긍하는 객관적인 수행능력 평가는 반드시 필수적이라고 할 수 있다. 이런 이유로 한국프로야구 투수들의 경기력 결과와 연봉 간의 패턴 분석은 중요한 의의를 가진다.

KBO 투수들의 경기력과 연봉과의 관계에 대한 선행연구들을 살펴보면 연봉에 중요한 영향을 미치는 여러 가지 독립변수를 이용하여 회귀분석을 실시한 Lee와 Kang (2001), 방어율, 경기출장, 승리, 패전, 세이브, 투구이닝, 피안타, 볼넷, 삼진, 실점, 자책점을 이용하여 주성분 회귀분석을 실시한 Kim (2002), 신경망분석, 의사결정나무분석, 회귀분석에 데이터마이닝 기법을 적용하여 KBO 선수들의 연봉에 관한 분석을 다룬 Oh와 Lee (2003), 세이버메트릭스 수치들인 WHIP 및 FIP 등의 기록을 이용하여 투수 자신이 통제할 수 있는 기록과 연봉과의 관계를 분석한 Kim (2013) 등이 있다. 한편 KBO 타

[1] (16890) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

자들의 경기력과 연봉과의 관계에 대한 선행연구들은 야구를 수리적으로 분석하는 방법론인 세이버메트릭스 지수를 계산하고 안타, 타점, 볼넷, 도루 등을 다시 혼합하여 연봉과의 중회귀분석을 실시하여 유의한 회귀계수를 도출한 Lee (2006), KBO 타자들에 대한 세이버메트릭스 지수 값을 이용하여 경기력과 연봉간의 패턴을 분석한 Seung과 Kang (2012) 등이 있으며, KBO 리그에 대한 최신연구로는 투수평가지표에 대한 Lee (2014)의 연구, KBO에 적당한 타자력 지수를 제안한 Hong 등 (2016), 한국 프로야구의 승률을 로지스틱 모형과 프로빗 모형을 통해서 추정한 Kim 등 (2016), 한국시리즈와 같은 우승결정방식에서 베르누이 시행과 독립성을 가정하는 경우에 상위 팀들이 우승할 확률을 추정하는 방법을 제안한 Cho (2016) 등이 있다.

한편 미국 프로야구 (Major League Baseball; MLB)인 경우는 투수들의 경기력과 연봉과의 관계에 대한 연구들을 많이 접할 수 있는 데, 1990년대부터 오늘날에 이르기까지 메이저 리그의 각 팀이 자유계약시대에 사용했던 연봉 메커니즘에 대한 통찰력을 제공하는 연구들이 많이 개발되었다. 일련의 연구들은 종속변수로 연봉이나 로그를 취한 연봉을 사용하고 투수 경기력으로 판단되는 다양한 독립변수들을 사용하여 모형을 만들었다 (Lackritz, 1990; Marburger, 1994; Hoaglin과 Velleman, 1995; Bollinger와 Hotchkiss, 2003; Hakes와 Sauer, 2006; Hills와 Gregory, 2014). 취급된 MLB 데이터는 시기가 다달라서 Lackritz (1990)는 1985 및 1986 시즌 자료, Marburger (1994)는 1991 및 1992 시즌 자료, Hoagin과 Velleman (1995)은 1986년의 연봉자료, Bollinger와 Hotchkiss (2003)는 1987년부터 1993년 시즌 자료, Hakes와 Sauer (2006)는 2000년부터 2004년 시즌자료, Hills와 Gregory (2014)는 2011년부터 2013년 사이의 시즌 자료를 각각 사용하였다.

그런데 지금까지 수행된 MLB 리그 연구는 특정시점에서 조사한 횡단연구 뿐만 아니라 시점을 달리하여 변화를 조사한 종단연구도 병행되었지만 KBO 리그 연구는 횡단연구가 대부분이다. 따라서 당해 연도의 같은 경기력에 대해서도 연봉은 선수에 따라 많은 차이가 발생하는 등 설명하기가 난감한 경우들이 많았다. 예를 들면 프로야구 첫해인 투수가 10승 및 방어율 2점을 기록했다고 10년 동안 연평균 10승 및 방어율 2점 투수와 같은 대접을 받을 수는 없다. 아마도 KBO에서 종단연구가 힘들었던 이유는 연구 시기가 야구통계의 접근성을 높일 수 있는 지속적인 통계 데이터베이스 구축 및 업데이트가 더디고 충실한 컨텐츠의 제공이 불가능한 시기였던 탓이라고 간주되는데, 최근 우리나라에서도 아직 MLB 만큼은 아니더라도 빠르게 프로야구에 대한 세밀한 데이터까지 제공하는 사이트들이 늘어나고 있다. 이런 이유로 과거에는 취급할 수 없었던 연봉에 관한 투수들의 경기력을 살펴보면 기존 연구의 한계점들을 보완할 수 있고, 투수들의 경기력과 연봉의 패턴 분석에 관한 연구를 좀 더 객관화할 수 있을 것이라 기대하며 한층 더 나아가서 구단이 선수의 연봉을 결정하는데 도움이 될 것으로 간주된다. 따라서 본 연구와 기존 연구의 차이는 기존 연구에서 배제된 여러 가지 종단연구의 요소를 가미하여 투수의 경기력과 연봉과의 관계를 살펴보았다는 점에 있다고 할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 데이터의 구성 및 본 연구에서 사용한 여러 가지 투수평가지표에 대한 설명과 기술통계 값을 살펴보았으며, 3절에서는 제안된 회귀모형과 결과를 설명하였다. 끝으로 4절에서는 본 연구의 결론을 제시하였다.

2. 연구방법

2.1. 데이터의 구성

데이터는 한국프로야구 사설기록 사이트인 스탯티즈 (`http://www.statiz.co.kr`)에 기록되어 있는 2010년부터 2015년 사이의 연봉데이터와 투수 기록 및 2016년 연봉 데이터를 이용하였으며 분석대상은 모두 698명의 경기결과이다. 데이터의 출처가 사설기록 사이트인 것은 한국프로야구 공식 홈페이지에는 연봉 자료가 없기 때문이다. 데이터의 가공 및 분석은 SAS university edition과 SPSS (ver. 23K)를

사용하고 그래프 작성은 R을 이용하였으며 통계분석기법으로는 상관분석 및 회귀분석을 활용하였다. 조사된 선수들은 KIA 77명, LG 87명, NC 28명, SK 83명, 넥센 67명, 두산 85명, 롯데 79명, 삼성 94명, 한화 79명, 그리고 KT 19명으로 삼성이 가장 많았으며, KT가 가장 적었다.

Table 2.1 Descriptive statistics for included variables

	MEAN	STD	MAX	MIN
G	26.28	20.196	80	1
ERA	5.8192	5.58340	81.00	0.00
FIP	5.2400	3.10291	36.86	-2.81
WHIP	1.6754	.90736	12.00	0.00
WAR	.9014	1.74627	22.00	-1.30
W	2.66	3.597	18	0
L	2.60	2.906	15	0
SV	1.31	4.722	47	0
IP	48.0699	44.41108	194.00	0.00
AGE	27.15	4.618	41	18
SALARY	11671.92	14570.976	125000	2400

서론에서 언급한 선행연구들을 참조하여 투수연봉에 영향을 줄 것으로 예상되어 고려한 설명변수들은 해당연도 (YEAR), 팀의 종류 (TEAM), 게임의 수 (G), 평균자책점 (ERA), 수비무관 평균자책점 (FIP), 이닝 당 안타 및 볼넷 허용률 (WHIP), 대체선수 대비 승리기여도 (WAR), 선발출장 게임수 (ST), 승 (W), 패 (L), 세이브 (SV), 투구이닝 수 (IP), 자유계약선수 여부 (FA), 출생년도 (BYEAR), 프로야구 시작연도 (SYEAR), 나이 (AGE), 경험연수 (EXP)이며 반응변수로는 선형회귀모형의 정규성 가정을 위하여 단위는 만원을 사용한 연봉 (SALARY)에 로그를 취한 로그연봉 (LOGSALARY)을 사용하였다. Table 2.1은 본 연구에 포함된 대표적인 투수 통계량에 대한 기술통계 값을 보여준다. 평균게임 수는 26.28, ERA와 FIP는 평균값은 비슷하나 표준편차, 최댓값 및 최솟값 모두 ERA의 경우가 FIP보다 약간 크며, 1이닝 당 평균적으로 1.68명 정도의 주자를 내보내며, 평균 WAR은 0.9, 승과 패의 평균은 각각 2.66과 2.60, 평균 세이브 수는 1.31, 평균 투구이닝 수는 48.07, 평균 나이는 27.15세, 평균 연봉은 11,672 (만원)으로 나타났다.

2.2. 투수지표

다양한 투수들의 평가지표 중 본 연구에서 사용한 보편화된 지표들은 다음과 같으며, 사용된 영어약자는 각각 자책점 (ER), 투구 이닝수 (IP), 삼진 (K), 볼넷 (BB), 피안타 (H), 피홈런 (HR)과 같다.

2.2.1. W-L

승리-패배 (win-loss record)는 투수를 평가하는 가장 대표적이고 오래된 지표로 많이 이길수록, 적게 질수록 우수한 투수가 된다. KBO 리그에서도 투수지표 중 가장 오래되고 많이 사용하는 평가기록이다. 투수 다승은 의미가 매우 큰 지표였지만 현대야구에서는 선발투수가 평균적으로 책임지는 투구이닝이 점점 더 낮아지고 있기 때문에 약간씩 중요도가 떨어지고 있다고 할 수 있다. Kim (2013)에 의하면 2012년 투수의 고과에 바탕을 둔 2013년 KBO 투수들의 연봉에 가장 높은 영향을 미치는 것은 승(W)으로 나타났을 정도로 여전히 투수기록의 핵심지표라고 할 수 있다.

2.2.2. ERA (Earned runs average)

평균자책점 (ERA)은 9이닝 당 투수 책임으로 허용한 점수인 자책점 (earned run; ER)의 비율을 의

미한다. ERA는 다승 (W)과 함께 가장 보편화된 고전적인 투수평가의 기준이지만 팀 전체의 책임인 실점을 투수만의 통계량으로 취급하는 모순이 있다.

$$ERA = 9(ER)/IP$$

2.2.3. WHIP (Walks plus hits divided by Innings pitched)

세이버메트리션들의 가장 대표적인 투수지표로 이닝 당 출루 허용율을 의미한다. 볼넷과 안타를 포함한 수치를 투구이닝으로 나눈 수치로 WHIP의 값이 적을수록 좋은 투수를 의미한다. 하지만 사구와 장타의 요소를 완전히 무시했다는 점에서 비판적인 시각도 많이 있다.

$$WHIP = (BB + H)/IP$$

2.2.4. FIP (Fielding independent pitching)

ERA 대체 통계로 수비무관 평균자책점이라는 뜻으로 실제로 투수가 통제할 수 있는 영역인 삼진, 볼넷, 홈런, 사구를 갖고 평균자책점의 형태로 나타낸 DIPS (defense independent pitching stats)를 기반으로 좀 더 실용적으로 사용할 수 있게 개량한 투수지표이다. 투수의 미래 성적을 예측하는데 유용하며 평균자책점과 같은 형태로 산출되기 때문에 비교해서 살펴보기가 편하다.

$$FIP = \frac{13(HR) + 3(BB) - 2(K)}{IP} + 3.20$$

2.2.5. WAR (Wins above replacement)

대체선수 대비 승리기여도로 설명되는 통계량으로 세이버메트릭스에서 선수의 가치를 평가하는 대표적인 지표이다. WAR의 값이 1인 의미는 대체선수에 비해 팀의 1승을 더 생산했다는 뜻으로 특정 선수가 보통 선수에 비해 팀의 승리에 기여도를 계산한 값이다. 이밖에도 투수의 평가지표로는 투수의 9이닝 동안 탈삼진 개수를 뜻하는 K/9, 9이닝 동안 허용한 볼넷의 개수를 뜻하는 BB/9, 볼넷 하나당 몇 개의 삼진을 잡았는지 보여주는 기록인 K/BB 등이 있으나 본 연구에서는 사용하지 않았는데, 이들 통계량은 선행연구에서도 연봉에 기여한다는 결과가 없었으며, 이들 통계량의 결과는 FIP 또는 WAR에 포함되었다고 할 수 있기 때문이다.

3. 분석 및 결과

본 연구에서는 로그연봉 값을 종속변수로 하였으며 투수 경기력 지표로 사용된 독립변수로는 특정연도의 수행평가지표인 W, ERA, WHIP, FIP, WAR과 세이브 수 (SV)와 같은 6개의 지표와 비수행 평가지표로는 나이 (AGE), 경험 (EXP), 선발투수 등판횟수 (ST), 자유계약 해당여부 (FA), 팀의 종류 (TEAM)를 사용하였다. 비수행 평가지표로 고려된 변수들에 대한 배경설명은 나이, 경험 등은 상식적으로 연봉에 효과를 줄 것으로 간주되었고, 투수자원이 한정되어 있는 KBO의 경우 선발투수의 평균 소화이닝이 중요하게 인식되고, 국내 FA시장에서 매년 역대 연봉협상 최고액을 갱신하는 계약들이 체결되는 등 FA 여부, 그리고 구단의 종류에 따라 연봉지급 예산규모가 다르고 연봉산출지표도 다르기 때문이다. 하지만 외국인 투수들은 연봉 산정에 주관적인 요소가 가미된 특수성이 있는 관계로 분석에서 제외하였다. Table 3.1은 본 연구에서 사용한 24개의 변수들과 의미를 보여준다. 변수의 정의를 좀 더 자세히 설명하면 연령 X_1은 해당연도-출생연도로 계산하고, 연령의 제곱 X_2는 나이가 많아지면 전성기를 지나게 되기 때문에 그런 가능성을 대비하기 위해 사용하였으며, 각 선수들의 경험연수를 의미하는

X_3는 현재연도-입단연도로 계산하고, 해당연도 X_4는 간편계산을 위하여 해당연도-2010으로 계산하였다. 한편 자유계약 여부를 설명하는 X_5는 해당연도 자유계약건이 있으면 1, 없으면 0으로 사용하였으며, X_{12}는 참여한 게임의 수, X_{13}은 세이브 수, X_{14}는 승률, X_{15}는 선발 출장한 게임의 수를 각각 의미하며, 변수 X_{16}부터 X_{24}까지는 팀의 종류를 나타내는 더미변수들이다.

Table 3.1 Description of predictors for regression model

Predictor	Meaning	Predictor	Meaning
X_1	age	X_{13}	saves
X_2	age^2	X_{14}	winning percentage
X_3	experience	X_{15}	games starting
X_4	year	X_{16}	team=n exen
X_5	free agency eligible	X_{17}	team=lotte
X_6	W	X_{18}	team=samsung
X_7	ERA	X_{19}	team=hanhwa
X_8	WHIP	X_{20}	team=KIA
X_9	FIP	X_{21}	team=KT
X_{10}	WAR	X_{22}	team=LG
X_{11}	IP	X_{23}	team=NC
X_{12}	G	X_{24}	team=SK

Table 3.2 Model summary for regression analysis

R Square	Adjusted R Square	Std. Error of the Estimate
0.747	0.742	0.44982

Table 3.3 Estimated regression model coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	6.938	.151		46.076	.000
X1	.036	.007	0.186	5.392	.000
X3	.024	.007	0.122	3.515	.000
X4	.050	.009	0.108	5.396	.000
X5	.455	.113	0.079	4.030	.000
X6	.068	.010	0.275	6.867	.000
X10	.024	.012	0.047	2.026	.043
X12	.014	.001	0.312	12.271	.000
X13	.030	.004	0.159	7.584	.000
X15	.028	.004	0.263	7.541	.000
X18	.179	.052	0.069	3.457	.001
X19	-.204	.055	-.073	-3.679	.000
X21	-.237	.109	-.044	-2.187	.029
X22	-.116	.053	-.043	-2.180	.030

회귀모형 설정은 단계선택법을 이용하였으며 그 결과가 Table 3.2와 Table 3.3이다. 선택된 회귀모형의 결정계수는 0.747로 나타났으며, 분산분석표는 지면 관계상 생략되었지만 p값은 $p < 0.001$로 회귀직선은 유의수준 1%에서 매우 유의한 것으로 나타났다. Table 3.3은 추정된 회귀식, 표준화 회귀계수를 보여주는데, 선택되어진 변수들은 모두 유의수준 5%에서 유의하였다. 하지만 y-축에 대한 이상치가 너무 많았고 따라서 모형의 개선은 충분히 가능하다고 판단되었다. y-축에 대한 이상치가 많은 이유는 대별해서 대략 2가지로 간주되는데, Figure 3.1과 Figure 3.2를 보면 알 수 있다.

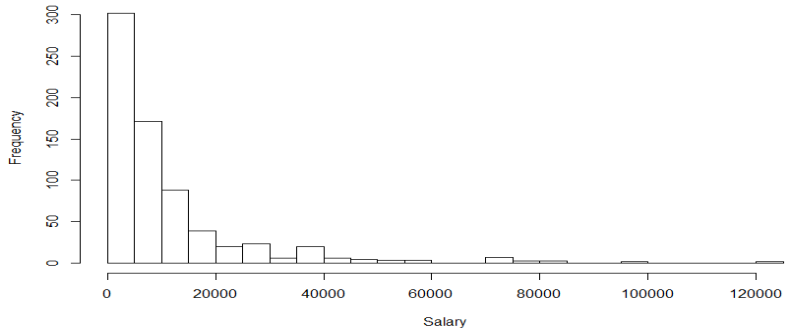


Figure 3.1 Histogram for salary

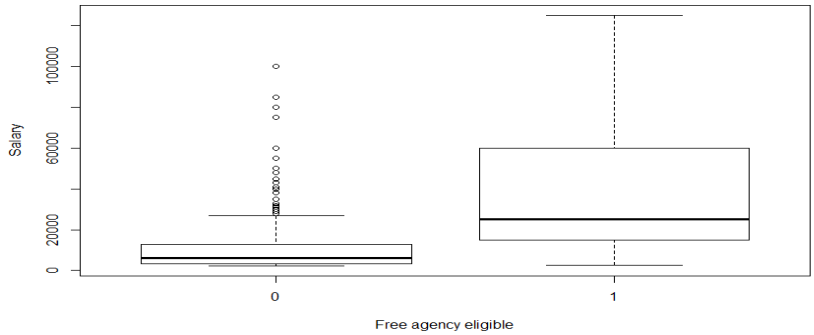


Figure 3.2 Box plot with two FA groups for salary

Figure 3.1은 연봉에 대한 히스토그램을 보여주는 데, 대부분이 저연봉으로 판단되며 또한 정규분포가 아닌 왜도가 매우 큰 값인 분포임을 알 수 있다. Figure 3.2는 FA 성사여부에 따른 연봉의 차이를 보여주는 데, FA가 있는 그룹이 없는 그룹에 비해 월등하게 연봉이 많다. 따라서 이런 이유로 y-축에 대한 이상치를 확인할 수 있는 스튜던트 제외잔차 (studentized deleted residual) r_i가 매우 크거나 작은 값이 많이 발생한다. 보통 선형회귀분석에서는 $|r_i| > 2$인 경우에 관측치를 이상치로 판단하는 데, 본 연구에서는 절단점을 2로 잡는 경우에 데이터의 1/3 이상을 제거하여야 했기 때문에 부득이하게 i번째 관측치의 $|r_i|$가 2.5 미만인 데이터만 분석에서 사용하였다. 그 결과 64개의 데이터가 삭제되고 634개의 자료를 이용하여 모형이 완성되었는데 그 결과 결정계수는 84.5%로 증가하고 모형도 변수 X_{21}과 X_{22}가 제거된 모형이 선택되어졌으며 자세한 모형은 Table 3.4와 같다.

Table 3.4를 살펴보면 분산팽창요인 (VIF)의 최대값이 4.269이므로 다중공선성 문제는 없다고 볼 수 있으며, 유의한 회귀계수로부터 KBO 투수들의 연봉은 매우 간단한 구조를 가진다고 설명할 수 있는데 나이 (X_1)가 많고, 경험 (X_3)이 많고, 연도 (X_4)가 많고, 자유계약 (X_5)이 있고, 승리의 수 (X_6)가 많고, WAR (X_{10})이 높고, 게임 수 (X_{12})가 많고, 세이브 수 (X_{13})가 많고, 선발게임 수 (X_{15})가 많고, 팀이 삼성 (X_{18})이면 많고, 팀이 한화 (X_{19})이면 적다고 해석되어진다. 물론 위의 결과 중에 팀의 종류에 종속되는 결과는 2010년부터 2015년까지의 결과이고 향후에는 어떤 변화가 있을 지는 각 구단의 경제적 여건 및 입장에 달렸을 것이다. 그런데 표준화회귀계수를 살펴보면 좀 더 중요도가 높은 변수가 무엇인지 알 수 있는데, 상위 1위부터 3위인 (X_{12}, X_6, X_{15}) 변수들은 모두 많은 게임을 치루고 선발투수이며 다승을 하는 요건으로서 이 사실은 투수자원이 부족한 우리나라인 경우 가장 뛰어난 투수들이 앞장

서서 다른 투수들 이상으로 이닝을 더 책임을 져야 운영이 가능하고 이런 이유로 선발 투수는 팀에서 가장 중요한 자산 중 하나라고 간주하는 것 같다.

Table 3.4 Estimated regression model coefficient for final model

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	VIF
	B	Std. Error	Beta			
(Constant)	6.931	.111		62.380	.000	
X1	.035	.005	.205	7.187	.000	3.259
X3	.013	.005	.075	2.620	.009	3.255
X4	.040	.007	.095	5.903	.000	1.043
X5	.576	.088	.106	6.561	.000	1.052
X6	.071	.007	.322	9.893	.000	4.269
X10	.020	.008	.045	2.354	.019	1.447
X12	.013	.001	.328	15.634	.000	1.766
X13	.030	.003	.178	10.317	.000	1.198
X15	.027	.003	.276	9.772	.000	3.215
X18	.192	.038	.082	5.072	.000	1.047
X19	-.169	.040	-.068	-4.223	.000	1.029

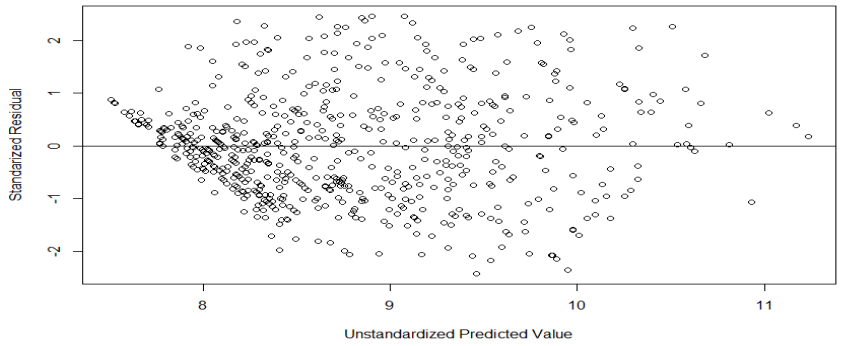
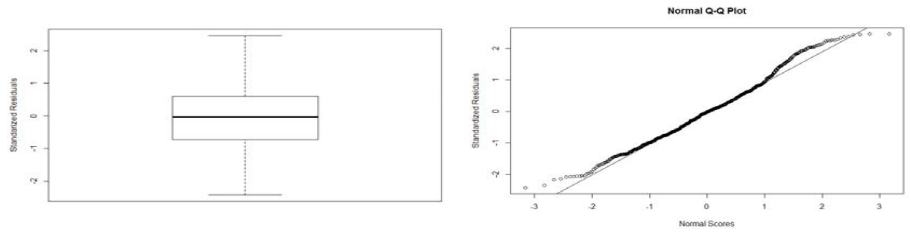


Figure 3.3 Residuals versus fits plot

한편 투수 자체의 능력이라고 볼 수 있는 ERA, WHIP, FIP, WAR 중에서는 WAR 만이 투수 연봉에 영향을 주었다. 이 사실은 누락된 통계량들이 연봉에 영향을 주지 않는다는 뜻이 아니라 이들 값들이 서로 상관관계가 매우 커서 다중공선성 문제로 변수선택 과정에서 배제된 것으로 판단되며, 투수능력을 판단하는 여러 가지 통계량 중에서 한국프로야구에서는 WAR이 가장 연봉과 크다고 판단되지만 선발투수에 관한 다른 통계량보다 현저하게 표준화회귀계수가 낮다. 즉, 투수 자체의 능력만으로 연봉이 결정되는 강도는 약하다고 볼 수밖에 없다.



(a) (b)

Figure 3.4 Box plot and QQ plot for residuals

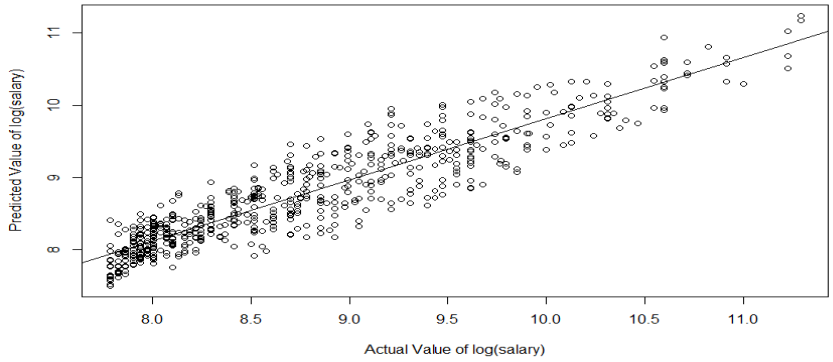


Figure 3.5 Scatterplot of predicted vs. actual values

Table 3.4의 식을 이용하여 작성한 잔차 도표가 Figure 3.3인데 패턴을 보아 등분산성을 가정할 수 있고, Figure 3.4는 잔차들에 대한 상자그림과 정규확률지그림을 보여주는 데, 오차들의 정규성을 의심할 필요가 없다고 판단되어진다. Figure 3.5는 실제의 로그연봉과 추정된 로그연봉의 관계를 보여주는 산점도이다. 대체적으로 실제 로그연봉을 추정된 로그연봉으로 잘 예측하고 있으며, 두 변수의 관계를 가장 잘 설명하는 회귀직선은 y를 실제 로그연봉, x를 추정된 로그연봉이라고 할 때 그림에도 나타나있지만 y=x로 판명되었다.

4. 결론

본 연구에서는 한국 프로야구 투수들의 수행능력과 비수행능력을 조사하여 연봉을 설명하는 선형회귀모형을 제안하였다. 수행능력을 평가하기 위해 사용한 측도로는 승리의 수, 평균자책점, WHIP, FIP, WAR, 패배의 수, 게임의 수, 세이브 수, 승률, 선발출장횟수를 고려하였으며, 비수행능력을 평가하기 위해 사용한 측도로는 나이, 경험, 연도, FA여부, 팀의 종류였다. 그 결과 투수들의 평가는 대부분 우수한 선발투수를 높게 평가하는 것으로 나타났다. 이점은 프로야구 초창기 때의 평가와 크게 달라진 점이 없다는 사실을 시사하며 논문 내용에는 생략되어 있지만 투수의 경기력으로 채택된 유일한 변수인 WAR의 표준화회귀계수 값도 연도에 따라 뚜렷하게 증가하는 패턴도 찾을 수가 없었다. 또한 최종모형에서 특이한 것은 투수가 얼마나 잘하는지 알아볼 수 있는 가장 기본적이며 대표적인 기록이 평균자책점 (ERA)인데 연봉에 영향을 주는 요인에 관한 회귀모형의 중요변수에서 빠진다는 점이다. 이 점에 대한 논쟁은 MLB에서도 마찬가지로 발생하는 데, Hills 등 (2014)에 의하면 여러 가지 논쟁에서 모두 ERA가 빠지는 사실을 확인할 수 있다.

프로야구 투수들은 당연하게 경기력이 높은 선수가 더 많은 연봉을 받는 것이 타당하기 때문에 앞으로의 투수 연봉산정은 투수 자체가 통제할 수 있는 기록들을 중심으로 이루어지는 것이 바람직하다. 왜냐하면 투수에 대한 세이버메트릭스 지수들은 투수 스스로 통제할 수 있는 기록을 가지고 뛰어난 투수를 과학적으로 판별하여 주기 때문이다. 본 연구의 목적은 우수한 경기력을 보였음에도 불구하고 불합리한 대우를 받는 선수들이 없기를 바라는 객관적인 연봉을 결정하는데 도움을 주고자 하는 데 있으며 나아가 투수 세이버메트릭스 측도를 좀 더 사용하여 보다 평가의 타당성을 갖춘 연봉 시스템이 완비되었으면 하는 희망이다. 본 연구에서는 다룰 수 없었던 구장 효과, 투수들의 인기도, 생활태도 등 다른 요인들도 투수들의 가치를 정하는 데 영향을 미치기 때문에 선수에 관련된 다양한 데이터를 손쉽게 구할 수 있다면 좀 더 프로야구 투수들의 경기력과 연봉의 패턴을 더욱 정확한 결과를 얻을 수 있을 것이며 이 부분은 향후 연구과제로 남겨둔다.

References

Bollinger, C. and Hotchkiss, J. (2003). The upside potential of hiring risky workers: Evidence from the baseball industry. *Journal of Labor Economics*, **21**, 923-944.

Cho, D. H. (2016). The winning probability in Korean series of Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **27**, 663-676.

Hakes, J. and Sauer, R. (2006). An economic evaluation of the moneyball hypothesis. *The Journal of Economic Perspectives*, **20**, 173-186.

Hills, C. and and Gregory, M. (2014). Professional baseball pitchers' performance and its effect on salary, https://www.overleaf.com/articles/professional-baseball-pitchers-performance-and-its-effect-on-salary/xndsqqqnrynm#share.

Hoaglin, D. and Velleman, P. (1995). A critical look at some analyses of major league baseball salaries. The *American Statistician*, **49**, 277-285.

Hong, C. S., Kim, J. Y. and Shin, D. S. (2016). Alternative hitting ability index for KBO. *Journal of the Korean Data & Information Science Society*, **27**, 677-687.

Kim, E. S. (2002). The relationship of game performance and annual salary for korean professional baseball pitchers. *Journal of Korean Sociology of Sport*,**15**, 95-104.

Kim, S. K. and Lee, Y. H. (2016). The estimation of winning rate in Korean professional baseball league. *Journal of the Korean Data & Information Science Society*, **27**, 653-661.

Kim, Y. H. (2013). *A study of determinants of Korean baseball pitchers salary*, Master Thesis, Sogang University, Seoul.

Lackritz, J. (1990). Salary evaluation for professional baseball players. *The American Statistician*, **44**, 4-8.

Lee, M. G. (2006). Relationship between performance ability of professional baseball batters and annual salary based on sabermetrics, Master Thesis,, Kookmin University, Seoul.

Lee, J. T. (2014). Pitching grade index in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **25**, 485-492.

Lee, J. Y. and Kang, H. M. (2001). The relationship between annual salary and performance of korean professional baseball pitchers. *Journal of Korean Sociology of Sport*, **14**, 115-124.

Marburger, D. (1994). Bargaining power and the structure of salaries in major league baseball. *Managerial and Decision Economics*, **15**, 433-441.

Oh, K. M. and Lee, J. T. (2003). A model study on salaries of korean pro baseball players using data mining. *Journal of Korean Sociology of Sport*, **16**, 295-309.

Seung, H. B. and Kang, K. H. (2012). A study on relationship between the performance of professional baseball players and annual salary. *Journal of the Korean Data & Information Science Society*, **23**, 285-298.

Journal of the Korean Data & Information Science Society
2017, **28**(2), 317–326

http://dx.doi.org/10.7465/jkdi.2017.28.2.317
한국데이터정보과학회지

Analysis of factors affecting Korean professional baseball pitcher salaries

Jang Taek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 11 February 2017, revised 19 March 2017, accepted 27 March 2017

Abstract

In this paper, we investigate the effects of performance and non-performance variables attributed to Korean professional baseball pitchers on annual salary by the records about pitchers between 2010 and 2016. We select the variables in reference to previous research related to this topic. The models are then estimated using linear regression model. For pitchers, age, experience in the league, year, eligibility for free agency, the number of wins, WAR, the number of innings pitched, the number of games, the number of saves, the number of games started, and type of baseball team have a statistically significant effect. Among the notable factors, affecting pitchers salaries are largely measure of starting pitchers. Pitcher sabermetrics indexes were poorly reflected on annual salary. The model presented here can be used to remove any unobjective salary differences for Korean professional baseball pitchers.

Keywords: Korean professional baseball, pitcher salary, regression model, salary difference, starting pitcher.

[1] Professor, Department of Applied Statistics, Dankook University, Gyeonggi-do 16890, Korea.
E-mail: jtlee@dankook.ac.kr