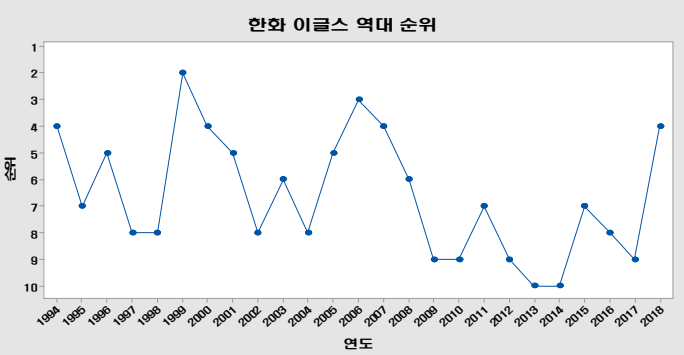
'Break The Frame'
세이버메트릭스로 본 한화 이글스의 비상

호서대학교
응용통계학과
이정호

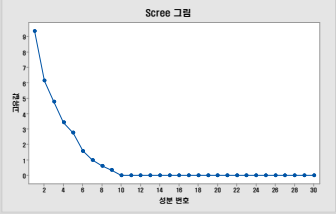
1. 배경 및 목적

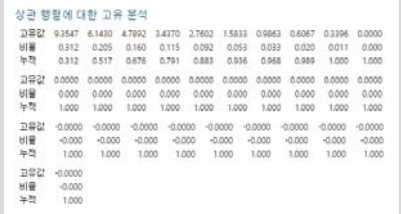
Break The Frame 2018년 한화 이글스의 팀 슬로건입니다. 한화 이글스가 2018년도 프로야구 판을 세차게 흔들고 있습니다. 전년도 8위 팀이 올해는 전반기 3위로 마무리 하였습니다. 어떤 이유에서 높은 승률을 가지게 된 지 확인해보려 합니다. 그리고 어떤 변수들이 승률에 영향을 끼치는지 알아보겠습니다. 변수는 KBO 공식 홈페이지에서 투수와 타자 각각 15개를 설정했습니다. 변수를 확인해보면 방어율과 실점 혹은 타석과 타수 등 변수들 값이 다중 공선성이 존재하게 됩니다. 그러므로 주성분 분석을 시행하여 변수축약을 한 뒤, 회귀분석을 시행하고 승률 예측 모형을 만들어서 승률 예측해보겠습니다.



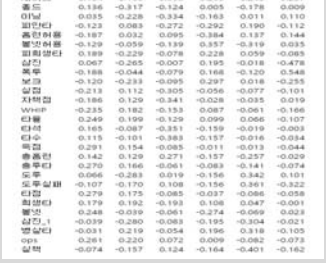
2. 주성분 분석

야구 데이터의 변수들은 측정단위가 서로 다르고, 값의 차이가 크기 때문에 표준화 변수를 사용해야 합니다. 그 후 주성분 분석을 실시합니다. 주성분 분석의 목적은 가장 적은 수의 주성분을 사용하여 분산의 최대량을 설명하는 것입니다. 변수의 척도가 서로 다르고 모든 변수에 동일한 가중치를 부여함으로써 올바른 결과를 얻기 위해 30개의 변수의 행렬 유형을 표준화된 상관계수로 선택하였습니다. 이 때 표준화란, 각 값에 평균을 빼준 뒤 표준편차로 나누어 주는 것 입니다.





주성분 분석을 통해서 나온 고유치와 누적 비율을 이용해서, 고유치 1 이상이고 누적 비율이 90% 이상인 주성분을 선택하였습니다. Scree 그림을 확인해보면 성분이 6-7으로 넘어갈 때 그래프가 급격이 변화합니다. 그러므로 성분 수를 6개로 선택하여 6개의 주성분을 얻었습니다.



주성분을 6개로 보았을 때 첫 번째 주성분에서 득점, 타점, 총 루타, OPS가 큰 양의 연관성이 있기 때문에 공격력이라 하겠습니다. 두 번째 주성분은 홀드, 삼진이 큰 음의 연관성이 있기 때문에 투수력이라 하겠습니다. 세 번째 주성분에서는 타수, 타석, 자책점, 이닝, 실점, 피안타가 큰 음의 연관성이 있기 때문에 체력이라 하겠습니다. 네 번째 주성분은 볼넷허용, 보크, 홈런허용, 피안타, 볼넷이 상관이 있기 때문에 제구력이라 하겠습니다. 다섯 번째 주성분은 도루실패, 도루가 큰 양의 상관이 있기 때문에 주력이라 하겠습니다. 마지막 주성분은 폭투, 삼진(투수), 보크가 높음으로 정신력이라 하겠습니다.

3. 변수 설명

투수	설명	타자	설명
방어율	게임당 투수가 내준 점수, 즉 투수 자책점의 평균이다.	타율	안타 수를 타수로 나눈 값
승	자신의 등판 리드 상태로 경기가 종료 되는 경우 자신은 승전 투수가 된다.	타석	타격 행위를 하기 위해 타석에 들어선 수
세이브	앞서고 있는 팀의 구원 투수가 경기 종료시까지 앞선 상황을 지키기 위한 투수의 기록이다.	타수	타석수에서 사사구, 희생타 등을 제외한 횟수
홀드	팀이 리드한 상황에서 중간 계투로 등판해 세이브 조건을 충족시키고 물러난 경우에 주어진다.	득점	주자가 득점권에 있을 경우
이닝	투구한 이닝수를 말한다.	총홈런	타자의 타구가 페어 라인 플대 안쪽의 외야 담장을 넘거나 혹은 수비 에러없이 모든 루를 다 밟았을 경우
피안타	투수가 투구한 공을 상대 타자가 쳐서 안타를 허용한 수이다.	총루타	단타는 1, 2루타는 2, 3루타는 3, 홈런은 4로 환산하여 모두 더한 것
피홈런	투수가 투구한 공을 상대 타자가 쳐서 홈런을 허용한 수이다.	도루	타격이나 볼넷, 몸에 맞는 볼과 상관없이 주자 자신의 힘으로 진루에 성공한 경우
볼넷	투수가 투구를 해서 볼넷 또는 몸에 맞는 볼을 허용한 수이다.	도루 실패	주자가 도루를 시도 했으나 태그 아웃된 경우
피희생타	희생번트와 희생플라이를 허용한 값이다.	타점	타격 행위로 인한 득점.
삼진	한 명의 타자를 상대로 한 타석에서 스트라이크 3개를 잡아 타자를 삼진 아웃 시키는 경우이다.	희생타	희생번트와 희생플라이를 더한 값
폭투	투수가 던진 공이 원바운드나 노바운드로 포수의 글러브에 들어가지 않는 경우를 의미한다.	볼넷	네 개의 볼을 뜻한다.
보크	부정 투구 동작으로 인한 투수의 반칙 행위를 의미한다.	삼진_타	타자가 3번째 스트라이크 판정을 받는 경우 삼진으로 아웃처리 됨
실점	투수의 책임이든 아니든 투수가 등판해 있을 때 상대에게 허용한 점수이다.	병살타	2명의 주자를 한 번의 수비 상황에서 아웃 처리 됨
자책점	야구 경기에서 투수의 책임이 되는 실점을 일컫는다.	OPS	출루율과 장타율을 더한 값
WHIP	이닝당 평균적으로 몇 명의 타자를 출루시켰는지를 뜻 함.	실책	야수가 처리할 수 있는 상황에서 공을 놓치는 등의 실수로 타자나 주자의 진루를 허용하는 경우를 말한다.

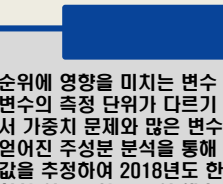
4. 회귀 분석

축약된 변수를 이용해서, 회귀분석을 하여 각 부분에 대한 적합 된 회귀모형을 만들고, y 값을 추정하여 순위를 예측해보겠습니다. 주성분 점수와 각 팀의 승률을 이용하여 회귀분석을 실시하기 위해서 먼저 각 주성분의 값에 표준화 값을 곱한 후 더하여 변수를 값을 얻었습니다. 적합 회귀모형을 실시하겠습니다.

반응 변수에 승률을 넣고 계량형 예측 변수에 각 주성분인 공격력-정신력까지의 변수를 입력합니다. 예측 값을 얻겠습니다. 적합치를 확인해보면 롯데와 KIA KT와 NC의 순위가 변한 것을 확인 할 수 있습니다. 따라서 적합치로 순위를 예측했을 때 한화 이글스는 3위 입니다.

모형이 얼마나 잘 접합되었는지 확인하기 해보겠습니다. R-제곱 값을 보겠습니다. R-제곱은 회귀분석의 설명력을 보여준다고 볼 수 있습니다. 이 모형은 반응 변동의 약 96%를

설명하고 있습니다. 분산분석 값을 확인해보면 회귀의 F-값에 P-값이 0.05 이내로 유의함을 보여주고 있습니다. 즉 회귀분석의 결과에 유의성이 있다고 말할 수 있습니다.



회귀분석 결과를 확인하겠습니다. 반응모형과 각 항의 간의 연관성을 통계적으로 유의한지를 확인하기 위해 p값을 유의수준과 비교하여 귀무가설을 평가해보겠습니다. 공격력을 제외하고 모두 값이 유의 하지 않음을 볼 수 있습니다. 이는 공격력을 제외한 모든 항이 연관성이 없다는 것을 볼 수 있습니다.

5. 결론

순위에 영향을 미치는 변수 30개를 사용하여 주성분 분석에 의한 회귀분석을 하였습니다. 변수의 측정 단위가 다르기 때문에 변수를 표준화하였습니다. 하지만 유사한 항목이 많아서 가중치 문제와 많은 변수에 의한 다중 공선성 문제에 의해서 주성분 분석을 하였습니다. 얻어진 주성분 분석을 통해 축약된 변수와 승률을 이용하여 적합 회귀 모형을 만들고, 값을 추정하여 2018년도 한국 프로야구 정규 시즌 순위를 비교하였습니다. 결과와 같이 한화 이글스의 승률이 예측값과 나와서 3위를 하여 10년 만에 가을 야구에 가서 플레이오프 한국시리즈 진출 나아가서 우승까지 노려봤으면 좋겠습니다. 한화 이글스 파이팅!