

2018 프로야구 승률 예측

YSABR

김조현 김철 박지원 현채연 홍진우



CONTENTS

- 01 단순 로지스틱 회귀(= 피타고리안 기대 승률)
- 02 다중 로지스틱 회귀
- 03 다중 로지스틱 회귀 + 신경망 모델



Intro

Sabermetrics

야구를 ‘통계학적 / 수학적’ 으로 분석하는 방법론



정규시즌 우승 기준

승률을 기준으로 순위를 정한다

$$\text{승률} = \text{승리 수} / (\text{승리 수} + \text{패배 수})$$



1

단순 로지스틱 (피타고리안 승률 예측)



Logistic Regression

$$\text{피타고리안 승률} = \frac{\text{총 득점}^\beta}{\text{총 득점}^\beta + \text{총 실점}^\beta} = \frac{1}{1 + (\text{총 실점} / \text{총 득점})^\beta}$$



Logistic Regression

$$\text{피타고리안 승률} = \frac{\text{총 득점}^\beta}{\text{총 득점}^\beta + \text{총 실점}^\beta} = \frac{1}{1 + (\text{총 실점} / \text{총 득점})^\beta}$$

$$\text{피타고리안 승률} = \frac{\text{총 득점}^\beta}{\text{총 득점}^\beta + \text{총 실점}^\beta}$$

$$\frac{\text{피타고리안 승률}}{1 - \text{피타고리안 승률}} = \left(\frac{\text{총 득점}}{\text{총 실점}} \right)^\beta$$

$$\log \left(\frac{\text{피타고리안 승률}}{1 - \text{피타고리안 승률}} \right) = \beta \log \left(\frac{\text{총 득점}}{\text{총 실점}} \right)$$

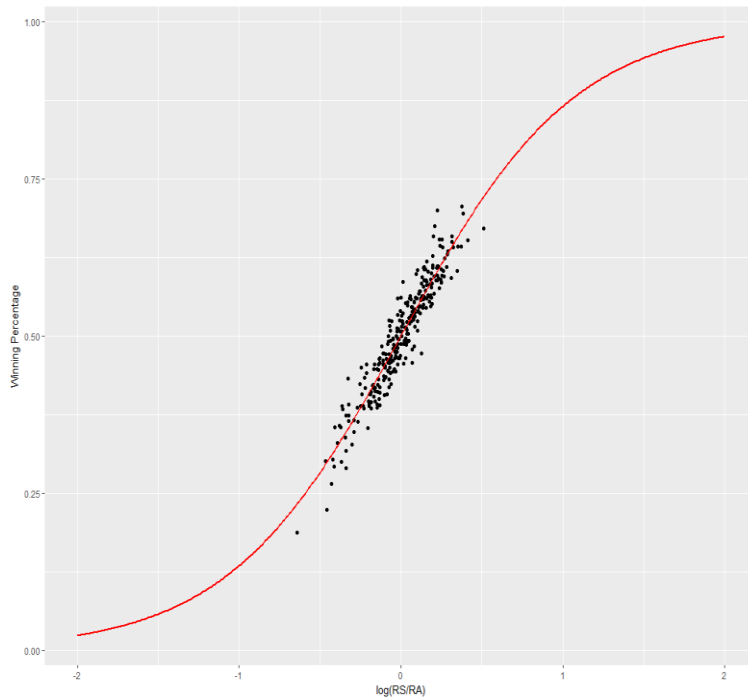
$$\log \left(\frac{\text{피타고리안 승률}}{1 - \text{피타고리안 승률}} \right) = \alpha + \beta \log \left(\frac{\text{총 득점}}{\text{총 실점}} \right)$$

→ 단순 로지스틱
회귀 모형



Logistic Regression

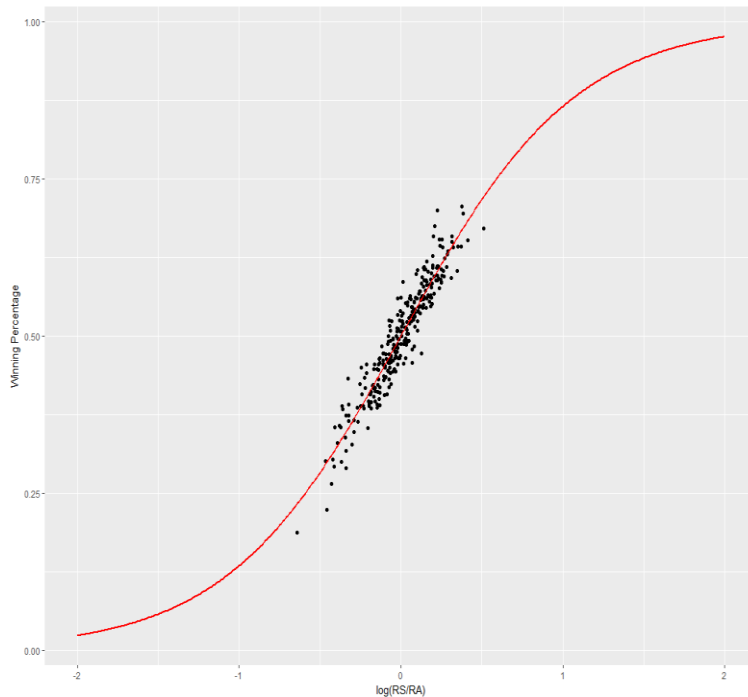
Coefficient	Estimate	Std.Error	Z-value	P-value
Intercept	0.001018	0.010743	0.095	0.925
$\log(\text{총 득점} / \text{총 실점})$	1.860938	0.060805	30.605	$< 2e-16$





Logistic Regression

Coefficient	Estimate	Std.Error	Z-value	P-value
Intercept	0.001018	0.010743	0.095	0.925
log(총 득점/ 총실점)	1.860938	0.060805	30.605	$< 2e-16$



* 실제 KBO에서 사용하는 β 값은 2,
MLB에서 사용하는 β 값은 1.85



Logistic Regression

$$\text{피타고리안 승률} = \frac{\text{총 득점}^\beta}{\text{총 득점}^\beta + \text{총 실점}^\beta} = \frac{1}{1 + (\text{총 실점} / \text{총 득점})^\beta}$$



Logistic Regression

우리의 목표 :

현재 승률의 예측 (**X**) / 시즌 마지막 승률의 예측 (**0**)



우리의 목표 :

현재 승률의 예측 (X) / 시즌 마지막 승률의 예측 (0)



현재 팀의 전력으로 시즌 마지막까지 가능한
'총 득점의 수' & '총 실점의 수' 예측
(by. 단순 선형 회귀 분석)



Logistic Regression

RMSE = 0.0330





각 팀의 승률에 대한 예측 오차의
평균은 약 3.3%





















$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$



Logistic Regression

팀	승률
 두산	0.667
 한화	0.595
 SK	0.560
 LG	0.557
 KIA	0.493
 넥센	0.488
 롯데	0.467
 삼성	0.438
 KT	0.390
 NC	0.350



팀	승률
 두산 	0.594
 LG  2	0.559
 SK 	0.541
 KIA  1	0.530
 넥센  1	0.517
 롯데  1	0.497
 한화  5	0.496
 삼성 	0.456
 KT 	0.444
 NC 	0.360



DATA

득점과 실점 데이터 말고
다른 데이터도 사용할 수 있지 않을까?



- 스탯티즈 (<http://www.statiz.co.kr/main.php>)
- 변수 : team, season, avg, oba, slg, OPS, wOBA, wRC+, WAR, ERA, FIP, WHIP, ERA+, FIP+, POSADJ, RAAwithADJ, WAAwithADJ
- 누적 지표 (x), 비율 지표와 가공 지표 (o)



2

다중 로지스틱



Multinomial Logistic Regression

team, season, avg, oba, slg, OPS, wOBA, wRC+, WAR, ERA, FIP, WHIP, ERA+, FIP+, POSADJ, RAAwithADJ, WAAwithADJ



경험적 판단

season, avg, oba, slg, OPS, wRC+, WAR, ERA, FIP, WHIP



Step-wise selection

avg, OPS, wRC+, WAR, ERA, FIP



Multinomial Logistic Regression

avg

타율

OPS

장타율

wRC+

타자의 타석당 득점 생산력을 평가하는 지표

WAR

선수의 '공격'과 '수비'를 종합적으로 평가할 수 있는 지표

ERA

투수의 9이닝당 평균자책점

FIP

수비 무관 평균자책점



Multinomial Logistic Regression

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.1251	0.0435	-2.87	0.0045
ERA	1	-0.0748	0.00907	-8.25	<.0001
FIP	1	0.0241	0.0118	2.05	0.0415
OPS	1	0.4092	0.1830	2.24	0.0265
WAR	1	0.00466	0.000718	6.50	<.0001
avg	1	1.0147	0.3617	2.81	0.0055
wRC_	1	0.00177	0.000572	3.09	0.0023



avg > OPS > ERA > FIP > WAR > wRC+ 순으로
승률에 영향을 많이 미침



Multinomial Logistic Regression


RMSE = 0.0320























각 팀의 승률에 대한 예측 오차의
평균은 약 3.2%



Multinomial Logistic Regression

팀	승률
 두산	0.667
 한화	0.595
 SK	0.560
 LG	0.557
 KIA	0.493
 넥센	0.488
 롯데	0.467
 삼성	0.438
 KT	0.390
 NC	0.350



팀	승률
 두산 	0.543
 LG  2	0.534
 SK 	0.508
 KIA  1	0.492
 넥센  1	0.461
 롯데  1	0.460
 한화  5	0.455
 KT  1	0.404
 삼성  1	0.387
 NC 	0.281



3

다중 로지스틱 + 신경망 모델



Multinomial Logistic Regression + Multi-layer Perceptron

다중 회귀

로지스틱 회귀

Gradient Boosting

라쏘 회귀

Random Forest

신경망 모델



Multinomial Logistic Regression + Multi-layer Perceptron

다중 회귀

로지스틱 회귀

Gradient Boosting

라쏘 회귀

Random Forest

신경망 모델



로지스틱 회귀

신경망 모델



Multinomial Logistic Regression + Multi-layer Perceptron

로지스틱 회귀

신경망 모델

로지스틱 회귀
+ 신경망 모델



Multinomial Logistic Regression + Multi-layer Perceptron

로지스틱 회귀

신경망 모델

로지스틱 회귀
+ 신경망 모델



로지스틱 회귀
+
신경망 모델 1
(2 Layer, 5 Node)
+
신경망 모델 2
(2 Layer, 6 Node)
+
신경망 모델 3
(2 Layer, 7 Node)



Multinomial Logistic Regression + Multi-layer Perceptron

RMSE = 0.0303



각 팀의 승률에 대한 예측 오차의
평균은 약 3.03%



Multinomial Logistic Regression + Multi-layer Perceptron

RMSE = 0.0303



각 팀의 승률에 대한 예측 오차의
평균은 약 3.03%









하지만,

어떤 변수가 승리에 강한 영향을 미치는지 해석하기 어렵고





















이 모델을 사용했을 때에는 개별 승률의 오차를 제시할 수 없음



Multinomial Logistic Regression + Multi-layer Perceptron

팀	승률
 두산	0.667
 한화	0.595
 SK	0.560
 LG	0.557
 KIA	0.493
 넥센	0.488
 롯데	0.467
 삼성	0.438
 KT	0.390
 NC	0.350













팀	승률
 두산 	0.543
 LG  2	0.528
 SK 	0.489
 KIA  1	0.481
 롯데  2	0.458
 넥센 	0.448
 한화  5	0.447
 KT  1	0.394
 삼성  1	0.382
 NC 	0.260



RESULTS











단순 로지스틱

RMSE = 0.0330

팀	승률
 두산	0.594
 LG	0.559
 SK	0.541
 KIA	0.530
 넥센	0.517
 롯데	0.497
 한화	0.496
 삼성	0.456
 KT	0.444
 NC	0.360











다중 로지스틱

RMSE = 0.0320

팀	승률
 두산	0.543
 LG	0.534
 SK	0.508
 KIA	0.492
 넥센	0.461
 롯데	0.460
 한화	0.455
 KT	0.404
 삼성	0.387
 NC	0.281

다중 로지스틱 + 신경망 모델

RMSE = 0.0303

팀	승률
 두산	0.543
 LG	0.528
 SK	0.489
 KIA	0.481
 롯데	0.458
 넥센	0.448
 한화	0.447
 KT	0.394
 삼성	0.382
 NC	0.260



DEPTH?

덱스(선수층) 때문일까?

-> 비주전 선수들이 주전 선수들만큼의 기량을 가지고 있으면 '덱스가 두껍다' 고 함



DEPTH?

덱스(선수층) 때문일까?

-> 비주전 선수들이 주전 선수들만큼의 기량을 가지고 있으면 '덱스가 두껍다' 고 함

-> 내야 / 외야 / 포수로 나누어 각 팀의 주전들끼리의 WAR, 비주전끼리의 WAR 비교



DEPTH?

덱스(선수층) 때문일까?

-> 비주전 선수들이 주전 선수들만큼의 기량을 가지고 있으면 '덱스가 두껍다' 고 함

-> 내야 / 외야 / 포수로 나누어 각 팀의 주전들끼리의 WAR, 비주전끼리의 WAR 비교



두산 > KIA > LG > SK > 삼성 > 넥센 > KT > 한화 = 롯데 > NC

A photograph of a baseball field with a white baseball and red stitching resting on the brown dirt near a white chalk line. The text "Thank You" is overlaid in a large, bold, black font. The background shows a blurred view of the field and a green fence under a clear sky.

**Thank
You**