Journal of the Korean Data & Information Science Society
2016, **27**(3), 653–661

http://dx.doi.org/10.7465/jkdi.2016.27.3.653
한국데이터정보과학회지

한국 프로야구의 승률 추정

김순귀[1] · 이영훈[2]

[12]강릉원주대학교 정보통계학과

접수 2016년 2월 22일, 수정 2016년 3월 23일, 게재확정 2016년 3월 31일

요 약

본 연구에서는 한국 프로야구의 승률을 추정하기 위하여 야구 경기의 피타고라스 정리라고 불리우는 방법을 사용하였고, 이 방법을 확장한 일반화 피타고라스 정리도 이용하면서 일반화 피타고라스 정리의 최적 지수 값을 찾아보았다. 그리고 다른 추정 방법들인 로지스틱 모형과 프로빗 모형의 사용을 제안하였다. 평균제곱오차의 제곱근 (RMSE)을 판정기준으로, 피타고라스 정리와 제안된 모형들의 효율성을 서로 비교하였다. 사용한 자료는 1982년부터 2015년 7월까지의 모든 한국 프로야구 기록이며, 제안한 방법은 일반화 피타고라스 정리를 이용한 승률 추정 방법보다 평균제곱오차의 관점에서 다소 나아졌음을 보여준다.

주요용어: 로지스틱 모형, 일반화 피타고라스 정리, 집락분석, 프로빗 모형.

1. 서론

야구 경기에서 득점 (runs scored; rs)과 실점 (runs allowed; ra)은 기본적으로 매 경기 27번의 기회를 통하여 공격력과 수비력, 에러 등 경기 중 발생하는 총체적인 결과의 산물이다. 야구 팬들은 투수의 방어율이나 타자의 타율, 출루율 등의 지표 뿐 아니라 이번 시즌에는 어느 팀이 우승할 지에 대하여 많은 관심을 가진다. 이에 관한 논문으로, Cho와 Cho (2005), Lee와 Kim (2006) 등이 있다.

이에 야구의 승률을 추정하기 위하여, James (1982)가 승률 w는 득점 (rs)의 제곱을 득점 (rs)의 제곱과 실점 (ra)의 제곱의 합으로 나눈

$$\widehat{w} = \frac{\mathrm{rs}^2}{\mathrm{rs}^2 + \mathrm{ra}^2} \tag{1.1}$$

로 추정할 수 있음을 제안하였고, 이를 야구 경기의 피타고라스 정리라고 불렀다 (James, 1982).

그는 실제 승률과 공식에 의한 승률 추정값의 차이를 보정하기 위하여, 연구를 통해 식 (1.1)을 일반화한 일반화 피타고라스 정리

$$\widehat{w} = \frac{\mathrm{rs}^\gamma}{\mathrm{rs}^\gamma + \mathrm{ra}^\gamma}$$

도 제안하였다. 이 때 지수 γ는 RMSE 등의 판정기준을 최소로 하는 값으로 결정되는데, 미국의 메이저 리그인 경우 지수 γ를 2에서 1.83으로 낮추어 승률을 추정하는 것이 바람직하다고 설명하였다.

[1] (25457) 강원도 강릉시 죽헌길 7 (지변동), 강릉원주대학교 정보통계학과, 교수.
[2] 교신저자: (25457) 강원도 강릉시 죽헌길 7 (지변동), 강릉원주대학교 정보통계학과, 교수.
E-mail: yhlee@gwnu.ac.kr

한국 프로야구의 매 시즌 별 팀 당 경기 수가 동일하지 않으므로, 본 연구에서는 시즌 별 팀 당 경기 수 G로 나눈 시즌 별 경기 당 득점 수 rsg와 경기 당 실점 수 rag를 변수 rs와 ra 대신 사용하였다.

2절에서 승률을 정의하고, 3.1절에서 rsg–rag와 승률 간의 상관분석 및 rsg/rag와 승률 간의 상관분석을 하였다. 3.2절에서 승률이 서로 다른 집락을 식별하기 위하여, 변수 rsg와 rag를 표준화하여 집락분석을 하였다. 3.3절에서 일반화 피타고라스 정리를 이용했을 때의 최적 지수 값을 계산하였다. 3.4절에서 변수 rsg와 rag를 이용하여, 로지스틱 모형과 프로빗 모형을 가정하여 승률을 추정하였다. 3.5절에서는 제시된 여러 모형에서의 RMSE를 비교하였다. 4절에서는 분석 결과를 바탕으로 결론을 도출하였다.

2. 승률의 정의

한국야구위원회(KBO)의 승률계산법은 다음과 같다.

Table 2.1 KBO system of measuring winning rates

year	KBO system of measuring winning rates
1982 ∼ 1986	(W+T*0.5)/G
1987 ∼ 1997	W/(W+L). Ties were excluded
1998 ∼ 2002	W. Winning rates were disregarded
2003 ∼ 2004	W/(W+L). Ties were excluded
2005 ∼ 2007	W/(W+L). Ties were excluded
2008	W/G. Ties were abolished
2009 ∼ 2010	W/G. Ties were reintroduced. tie=loss
2011∼ now	W/(W+L). Ties were excluded

(W : number of wins, L : number of losses, T : number of ties, G : total number of games)

KBO 승률계산법에 의한 승률계산이 연도마다 조금씩 차이가 있다. 따라서 저자들은 2011년부터 현재까지 한국 프로야구 승률계산법에서 정의한

$$w = \frac{\mathrm{W}}{\mathrm{W}+L}$$

를 사용하기로 한다. 여기에서 W는 매 시즌 승리한 경기 수, L은 매 시즌 패배한 경기 수를 각각 나타낸다.

본 연구에서는 제안될 모형들의 효율성을 서로 비교하기 위하여, 일반적으로 많이 사용하는 평균제곱오차의 제곱근 (root mean square error; RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(w_i - \widehat{w}_i)^2}{N}}$$

을 사용하는데, 여기에서 w_i는 한국 프로야구 승률 계산법에서 정의한 i번 째 팀의 승률을, $\widehat{w}_i$는 피타고라스 정리 또는 제안한 모형을 이용한 승률의 추정값, N은 총 게임 수를 나타낸다. RMSE가 가장 작은 것이 제일 좋은 추정량이라 할 수 있다.

3. 한국 프로야구 자료의 분석

한국 프로야구 원년인 1982년부터 2014년까지의 시즌 별 각 팀에 대한 자료가 Table 3.1과 같으므로, 따라서 총 $N = 253$ 팀이다. 2015년 개막전부터 2015년 7월 31일까지의 자료 $n = 10$개 팀은 검정용 자료로 사용한다.

Table 3.1 Data of teams and games

year	number of teams of the season	G : number of games per team
1982 ∼ 1985	6	80 ∼ 110
1986 ∼ 1990	7	108 ∼ 120
1991 ∼ 2012	8	126 ∼ 133
2013 ∼ 2014	9	128
2015 ∼ now	10	128

본 연구에서는 변수 rsg와 rag를 이용하여 승률을 추정하고자 한다. 먼저 승률과 변수 rsg–rag (rsg와 rag의 차이), rsg/rag (rsg와 rag의 비율) 사이의 상관관계를 조사하고, 변수 rsg와 rag를 이용한 집락분석을 실시하여 어떤 집락이 승률이 좋은지 분석할 것이다. 다음에 James (1982)가 제안한 피타고라스 정리와 일반화 피타고라스 정리를 이용하여 승률을 추정하고, 마지막으로 로지스틱 모형과 프로빗 모형을 각각 적합시켜 James가 제시한 승률 추정법과 비교하고자 한다. 사실 이 논문에서 다루는 자료의 구조는 iid는 아니지만, 자료의 수가 $N = 253$으로 그리 작은 편은 아니므로 iid를 가정하여 기존 분석 방법을 사용하였다.

3.1. 상관분석

변수 rsg와 rag의 차이인 rsg–rag와 한국 프로야구 승률 계산법에서 정의한 승률 간의 상관분석을 한 결과와 산점도가 각각 Table 3.2와 Figure 3.1에 주어졌다. (11번 째 관측값을 이상값으로 식별하여, 상관분석의 경우에 한하여 제거하였다.) 참고로 변수 rsg/rag와 승률 간의 상관분석을 한 결과와 산점도가 각각 Table 3.2와 Figure 3.2에 있다.

Table 3.2 Correlation coefficients of winning rate and rsg–rag & rsg/rag

		rsg–rag	rsg/rag
winning rate	Pearson correlation coefficient	0.923	0.914
	p value	<0.001	<0.001

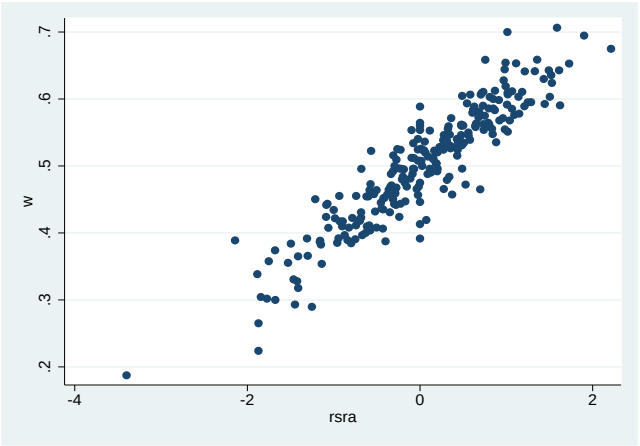

Figure 3.1 Scatter plot of winning rate & rsg–rag

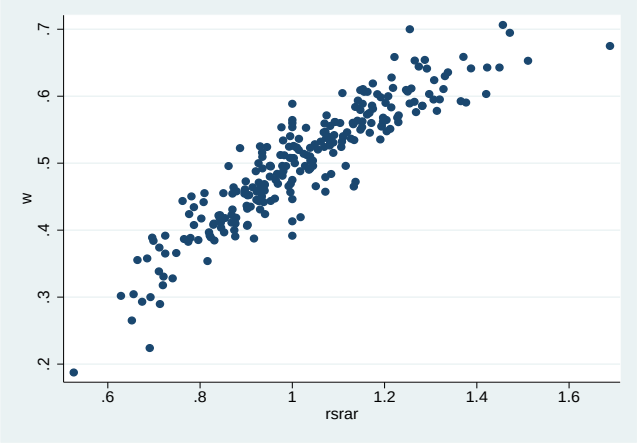

Figure 3.2 Scatter plot of winning rate & rsg/rag

위 결과에 의하면 rsg–rag와 승률 간의 상관계수의 값은 r=0.923 ($p < 0.001$)으로, rsg–rag의 값이 커질수록 승률이 높아지는 양의 상관관계를 보였다. 마찬가지로 rsg/rag와 승률 간의 상관계수의 값은 r=0.914 ($p < 0.001$)으로, rsg/rag의 값이 커질수록 승률이 높아지는 양의 상관관계가 나타났다.

3.2. 집락분석

승률이 서로 다른 집락을 식별하기 위하여, 변수 rsg와 rag를 표준화하여 집락분석을 하였다. 계층적 집락분석 중에서도 집락 내 거리의 오차제곱합을 최소로 하는 Ward의 방법을 선택하였고, 집락의 수는 4개로 하였다 (Kim과 Jhun, 1994 ; Huh, 2000). 결과를 보여 주는 dendrogram과 Table 3.3의 cluster history에 의하면, R^2이 급격히 감소하는 지점인 집락의 수를 3 또는 6으로 함이 타당한 듯 하지만, 그 절충안으로 집락의 수를 4개로 하여 해석하기 쉽게 하였고 또한 4개의 집락에서의 승률에 대한 F-검정이 유의한 결과를 얻었기 때문이다. 4개의 집락을 가지는 집락분석을 한 결과가 Table 3.4에, 4개의 집락을 나타내는 그래프가 Figure 3.3에 있다.

Table 3.3 Cluster History

Number of Clusters	Clusters	Joined	Freq	Semipartial R-Square	R-Square Tie
10	CL22	CL25	17	0.0182	.854
9	CL15	CL21	50	0.0212	.833
8	CL16	11	28	0.0214	.811
7	CL14	CL11	65	0.0254	.786
6	CL17	CL8	61	0.0282	.758
5	CL9	CL12	78	0.0667	.691
4	CL5	CL10	95	0.0701	.621
3	CL13	CL6	93	0.0772	.544
2	CL3	CL7	158	0.1558	.388
1	CL2	CL4	253	0.3879	.000

Table 3.4 Cluster analysis

	cluster 1	cluster 2	cluster 3	cluster 4
mean of rsg	3.956	5.162	4.116	5.059
mean of rag	4.745	5.110	3.807	3.894
number of data in cluster	65	95	61	32

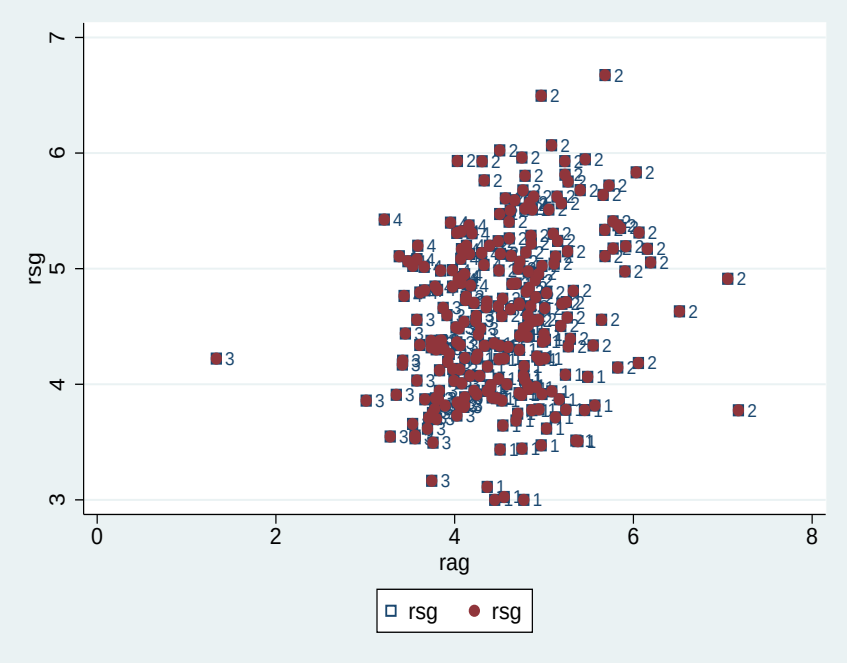

Figure 3.3 Scatter plot of rsg & rag (The natural numbers mean clusters.)

위 결과에 의하면 집락 4가 승률이 가장 높은 집단으로, 역시 rsg의 값이 평균보다 큰 값을 가지고 rag의 값은 매우 작았다. 집락 1은 승률이 가장 낮은 집단으로, 예상한 대로 rsg의 값이 제일 작았고 반면에 rag는 비교적 큰 값을 가졌다. 흥미있는 결과로 집락 2와 집락 3을 비교하면 승률이 비슷할 것으로 예상되었지만, 예상과 달리 집락 3의 승률이 집락 2의 승률보다 0.033 정도 높게 나왔다. 이 사실은 다음을 의미한다. 팀이 승리하기 위하여는, 득점을 적게 하더라도 실점을 최소화하여야 함을 강력하게 암시하고 있다. 이는 득점을 많이 하는 것보다는 실점을 적게 하는 것이 승리의 요인이 됨을 의미한다.

집락 별 승률을 비교한 분산분석의 결과가 Table 3.5에 있다.

Table 3.5 Comparison of winning rates of clusters

cluster	mean	standard error	test (Bonferroni & Scheffe)
1	0.417	0.0080	
2	0.499	0.0085	$4 > 3 > 2 > 1$
3	0.533	0.0065	
4	0.612	0.0070	

$F = 67.52$, $p < 0.001$

분산분석 결과 $p < 0.001$로 집락 간의 유의한 차이를 보여주었고, 사후분석을 한 결과 변수 rsg와 rag의 값이 큰 집락보다 상대적으로 작은 집락의 승률이 더 높음을 알 수 있었다. 즉 공격력이 좋은 팀보다는 투수력과 수비력을 겸비하여 실점을 적게 하는 팀의 승률이 더 높음을 확인할 수 있었다.

3.3. 피타고라스 정리

일반화 피타고라스 정리를 이용한 최적 지수 값은 1.71 정도로 계산되었는데 (N=253 팀의 자료를 이용), 이는 미국 메이저 리그인 경우 총득점과 총실점의 지수 값 1.83에 비해 다소 낮은 값으로 한국 프로야구의 승률을 추정할 수 있다는 의미이다 (Lee, 2015).

참고로 1982년부터 2005년까지의 한국 프로야구 기록을 이용한 최적 지수 값은 1.87, RMSE는 0.03095 (3.095%)로 계산되었다 (Lee, 2014 ; Lee와 Kim, 2006).

3.4. 로지스틱 모형과 프로빗 모형

변수 rsg와 rag를 이용하여, 로지스틱 모형을 가정하여 추정한 결과는 다음과 같다.

$$\widehat{w} = \frac{\exp(0.0152 + 0.3781\text{rsg} - 0.3848\text{rag})}{1 + \exp(0.0152 + 0.3781\text{rsg} - 0.3848\text{rag})}$$

여기에서 $\widehat{w}$는 승률의 추정값을 나타낸다. 이 때 적합한 모형의 카이제곱 값 $\chi^2(2)$ (독립변수의 수 = 2 이므로, 자유도는 2임)와 p값은 각각

$$\chi^2(2) = 786.81$$
$$p < 0.001$$

이므로, 두 변수 모두 승률에 유의한 영향을 주고 있음에 틀림이 없다.

마찬가지로 변수 rsg와 rag를 이용하여, 프로빗 모형을 가정하여 추정한 결과는 다음과 같다.

$$\widehat{w} = \Phi(0.00294 + 0.2358\text{rsg} - 0.2385\text{rag})$$

이 때 적합한 모형의 카이제곱 값 $\chi^2(2)$과 p값은 각각

$$\chi^2(2) = 785.73$$
$$p < 0.001$$

이므로, 두 변수 모두 승률에 유의한 영향을 주고 있음에 틀림이 없다 (Hosmer와 Lemeshow, 2000; Kim, 2014)

3.5. 여러 모형에서의 RMSE 비교

피타고라스 정리, 일반화 피타고라스 정리, 로지스틱 모형, 그리고 프로빗 모형을 가정하여 각각 승률을 추정하고, 여러 모형에서의 평균제곱오차의 제곱근 RMSE를 구한 것이 Table 3.6에 있다.

Table 3.6 Estimation methods and RMSE

	Pythagorean	generalized Pythagorean	logistic	probit
RMSE	0.0453	0.0437	0.0394	0.0395

Table 3.6에 의하면, 피타고라스 정리보다는 최적 지수 값 1.71인 일반화 피타고라스 정리를 이용하여 승률을 추정함이 RMSE 기준에서 다소 나아졌지만, 로지스틱 모형이나 프로빗 모형을 이용하여 승률을 설명함이 가장 적합성이 좋은 것으로 나타났다.

참고로 모형의 확인을 위하여, 새로운 자료에서 설명변수들의 특정한 값을 모형에 대입했을 때의 예측값과 새로운 자료의 반응값의 차이를 관찰하였다. 이를 위하여 MSPR (mean squared prediction error)

$$\text{MSPR} = \frac{\sum_{i=1}^{n}(w_i - \widehat{w}_i)^2}{n}$$

를 이용하는데, 여기에서 w_i는 새로운 자료의 관측값, $\widehat{w}_i$는 가정한 모형을 이용한 승률의 예측값, n은 자료의 수를 각각 나타낸다.

피타고라스 정리와 로지스틱 모형 등을 가정하여 MSPR과 MSPR의 제곱근을 비교한 것이 Table 3.7이다.

Table 3.7 MSPR

	Pythagorean	generalized Pythagorean	logistic	probit
MSPR	0.00075	0.00061	0.00084	0.00083
$\sqrt{\text{MSPR}}$	0.02739	0.02470	0.02898	0.02881

Table 3.7에 의하면, 일반화 피타고라스 정리를 이용한 승률의 추정이 다른 방법보다 다소 나은 값을 보여주지만, 그 차이가 미미하다고 볼 수 있다. 이는 모형의 확인을 위한 팀 수 ($n = 10$)가 적은 데에서 기인한 것으로 보인다. 따라서 한국 프로야구에서 각 팀의 승률을 추정하기 위하여, 피타고라스 정리와 일반화 피타고라스 정리 뿐 아니라 로지스틱 모형, 프로빗 모형 등의 다양한 모형을 사용하여 승률을 추정하고 예측하기를 제언한다.

4. 결론

본 연구에서는 rsg–rag, rsg/rag와 승률 간의 상관분석을 실시하였고, 변수 rsg와 rag를 이용하여 집락분석을 시도하였다. 그 결과 rsg–rag, rsg/rag의 값이 클수록 승률이 높아지는 경향이 있었고, 또한 변수 rsg와 rag의 값이 다소 높은 집락보다 두 변수의 값이 다소 낮은 집락의 승률이 더 높은 것으로 나타났다. 이 사실에 의하면, 한국 프로야구에서는 득점을 적게 하더라도 실점을 적게 하는 팀의 승률이 득점과 실점 모두 다소 높은 집락의 승률보다 더 높음을 의미한다. 더 나아가 변수 rsg와 rag를 사용하여 Bill James의 피타고라스 정리와 일반화 피타고라스 정리를 한국 프로야구에 적용하여 팀의 승률을 추정하였고, 마지막으로 rsg와 rag를 이용하여 로지스틱 모형과 프로빗 모형을 가정하여 승률을 추정하였다.

이렇게 유도된 식들은 모두 한국 프로야구에서의 승률을 잘 설명하고 있지만, 이 논문에서 제안한 로지스틱 모형과 프로빗 모형을 이용한 승률 추정이 피타고라스 정리를 이용한 것보다 다소 나은 것으로 나타났다.

야구 경기에서 승리하려면, 당연히 실점보다 득점이 많아야 함은 기본이다. 따라서 득점과 실점으로 승리가 결정되는 다른 종목 경기에서의 승률을 추정하기 위하여, 향후 일반화 피타고라스 정리를 적용하여 최적 지수 값을 계산하거나 로지스틱 모형 등을 이용하는 연구도 의미있을 것으로 생각된다.

References

Cho, Y. S. and Cho, Y. J. (2005). A study on winning percentage using batter's runs and pitcher's runs in Korean professional baseball league. *Journal of the Korean Data Analysis Society*, **7**, 2303-2312.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*, 2nd ed., Wiley, New York.
Huh, M. H. (2000). *Multivariate data analysis*, Freedom Academy, Seoul.
James, B. (1982). *The Bill James baseball abstract*, Ballantine Books, New York.
Kim, K. Y. and Jhun, M. S. (1994). *SAS cluster analysis*, Freedom Academy, Seoul.
Kim, S. K. (2014). *Understanding of logistic regression model*, Kyowoosa, Seoul.
Lee, J. T. (2014). Estimation of exponent value for Pythagorean method in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **25**, 493-499.
Lee, J. T. (2015). Measuring the accuracy of the Pythagorean theorem in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **26**, 653-659.
Lee, J. T. and Kim, Y. T. (2006). A study on the estimation of winning percentage in Korean pro-baseball. *Journal of the Korean Data Analysis Society*, **8**, 857-869.

Journal of the Korean Data & Information Science Society
2016, **27**(3), 653–661

http://dx.doi.org/10.7465/jkdi.2016.27.3.653
한국데이터정보과학회지

The estimation of winning rate in Korean professional baseball league

Soon-Kwi Kim[1] · Young-Hoon Lee[2]

[12]Department of Information Statistics, Gangneung-Wonju National University

Received 22 February 2016, revised 23 March 2016, accepted 31 March 2016

Abstract

In this paper, we provide a suitable optimal exponent in the generalized Pythagorean theorem and propose to use the logistic model & the probit model to estimate the winning rate in Korean professional baseball league. Under a criterion of root-mean-square-error (RMSE), the efficiencies of the proposed models have been compared with those of the Pythagorean theorem. We use the team historic win-loss records of Korean professional baseball league from 1982 to the first half of 2015, and the proposed methods show slight outperformances over the generalized Pythagorean method under the criterion of RMSE.

Keywords: Cluster analysis, generalized Pythagorean theorem, logistic model, probit model.

[1] Professor, Department of Information Statistics, Gangneung-Wonju National University, Gangneung 25457, Korea.
[2] Corresponding author: Professor, Department of Information Statistics, Gangneung-Wonju National University, Gangneung 25457, Korea. E-mail: yhlee@gwnu.ac.kr