DATA ANALYSIS REPORT

CJHSTAT CONSULT

COMMUNICATION

INTUITION

ACCURACY

연구기간 2020-05-31 ~ 2020-06-04

본 견본에는 실제 분석내용은 삭제되어 있습니다

연구자 응용통계학 석사 장 재 호[1)]

1) cjhsfl@gmail.com

연구주제

번호	항목	체크	비고
1	KLPGA와 LPGA의 상금, 평균타수 간 차이가 존재한다.	v	분석결과 문제 없음.
2	KLPGA 시즌결과물의 (최대)결정요인 탐색	v	모형 진단 후 재적합.
3	LPGA 시즌결과물의 (최대)결정요인 탐색	v	모형 진단 결과, 재적합 필요 없음.
4	KLPGA + LPGA 시즌결과물의 (최대)결정요인 탐색	v	모형 진단 후 재적합.
5	우승경력 有 선수들의 시즌결과물(KLPGA, LPGA) (최대)결정요인 탐색	v	(특이사항) KLPGA의 경우 birdies만 유의미한 변수로 선택됨.
6	우승경력 無 선수들의 시즌결과물(KLPGA, LPGA) (최대)결정요인 탐색		
7	우승경력 無 top10에 선수 중 향후 우승 가능성 예측 + 경기력 분석	v	차기 우승자 예측

[연구주제 요약] 대상기간은 2017 ~ 2019년임.

1. 기초통계 요약

가. 빈도분석

1) 결측치 분석

빈도분석 결과

이 검색되었다.

		year	rank	name	prize	scoring_avg	wins	plays	rookie_yr
N	유효								
	결측								
		top10finish_percentile	driving_accuracy	driving_distance	green_in_regulation	putting_avg	birdies	nationality	association
N	유효								
	결측								

가) **부적절한 코딩**

국적이 다음과 같이 같은 국적이 다르게 코딩되어 있는 경우, 잘못 코딩된 경우를 수정했다.

nationality	빈도	nationality	빈도
South Africa		KOREA	
australia		Mexico	
Australia		Newzealand	
canada		Norway	
Canada		Russia	
china		spain	
China		Spain	
Denmark		Swden	
England		Sweden	
France		Thailand	
Germany		U.S.A	
Japan		전체	
korea			
Korea			

나. 기술통계

본 분석에 앞서, 주어진 데이터의 구조는 선수, 연도, 협회 조합에 따른 총 360여건의 데이터이다. 따라서 같은 선수의 서로 다른 연도에 측정된 기록은 시계열 변수(time-series variable)이면서 자기상관성(autocorrelation)을 갖게 된다. 자기상관성이 존재할 경우 각 종속변수 관측치가 서로 독립이다라는 통계모형의 이론적 가정을 심각하게 위배하기 때문에, 개인기록 및 시즌 결과물을 각 선수 및 협회별로 평균을 낸 값을 데이터로 사용하며, 앞으로 진행될 모든 분석에 동일하게 사용하도록 한다. 각 장마다 자기상관성이 없음을 Durbin-Watson test를 통해 통계적으로 검증하였다.

1) 협회 별 기술통계

다음으로 기술통계 요약이다. N은 유효 관측치 수, Mean은 평균, St. Dev.는 표준편차, Min은 최소값, Pctl(25, 75)는 1사분위수와 3사분위수, Max는 최대값을 의미한다. 연도에 대해서 평균을 낸 값들을 이용했기 때문에 데이터 크기는 총 360개에서 총 173개로 변환되었다.

KLPGA & LPGA

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
rank	174						
prize(억원)	174						
scoring_avg	173						
wins	173						
plays	173						
top10finish_percentile	173						
driving_accuracy	171						
driving_distance	171						
green_in_regulation	173						
putting_avg	173						
birdies	173						

KLPGA

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
rank	87						
prize(억원)	87						
scoring_avg	87						
wins	86						
plays	86						
top10finish_percentile	86						
driving_accuracy	85						
driving_distance	85						
green_in_regulation	87						
putting_avg	87						
birdies	87						

LPGA

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
rank	87						
prize(억원)	87						
scoring_avg	86						
wins	87						
plays	87						
top10finish_percentile	87						
driving_accuracy	86						
driving_distance	86						
green_in_regulation	86						
putting_avg	86						
birdies	86						

다. 상관분석

1) **상관분석 및 상관플랏**

변수들의 피어슨(Pearson) 상관관계를 분석하였다. 그림에서 각 셀 별로 파란색은 양의 상관관계, 빨간색은 음의 상관관계, 북동쪽 방향으로 그려진 타원은 양의 상관관계가 강할수록 북동쪽 대각선에 가까워진다. 반면 북서쪽 방향으로 그려진 타원은 음의 상관관계가 강할수록 북서쪽 대각선에 가까워진다. 아무런 도형이 그려지지 않은 셀은 상관관계가 유의수준 0.05에서 유의하지 않은 변수 조합이다. 분석결과 대체로
이 서로 양의 상관관계를 가진다(그림에서 왼쪽 위). 반면 이들과 음의 상관관계를 가지는 변수들은 가 있다(그림에서 오른쪽 위).
는 서로 양의 상관관계를 가진다(그림에서 오른쪽 아래). 상관도표는 대각선에 대해서 대칭이기 때문에 한쪽만 보면 된다.

2. t-test

가. 주제 1. KLPGA와 LPGA의 상금, 평균타수 차이

1) 집단통계량

집단통계량					
	association	N	평균	표준화 편차	표준오차 평균
prize(억원)	KLPGA	87			
	LPGA	87			
scoring_avg	KLPGA	87			
	LPGA	86			

2) **검정**

독립표본 검정								
		Levene의 등분산 검정		평균의 동일성에 대한 T 검정				
		F	유의확률	t	자유도	유의확률 (양측)	평균차이	표준오차 차이
prize(억원)	등분산 가정							
	이분산 가정							
scoring_avg	등분산 가정							
	이분산 가정							

우선 prize에 대한 t 검정은 유의수준 0.05에서 등분산 검정을 기각했으므로, 이분산 가정 t-검정을 수행했다. 는 유의수준 0.05에서 등분산 가정을 채택했다. 검정 결과 모두 KLPGA 와 LPGA 간 유의미한 차이가 존재했다. 상금은 LPGA가 KLPGA에 비해 평균적으로 약 6.4억원 정도 액수가 컸으며, 평균타수는 KLPGA가 평균적으로 1만큼 더 컸다.

3. 전진선택법(forward-selection)을 이용한 선형회귀분석

가. 주제 2. KLPGA 시즌결과물(rank)의 (최대)결정요인 탐색

이 장에서는 Durbin-Watson 검정을 통해 회귀모형의 자기상관성이 사라졌음을 OLS의 모형진단에서 후술한다.

1) 전진선택

종속변수는 rank, prize 두 개로 나누어서 분석하였다. 모형에 대한 해설은 rank만 하기로 한다. rank의 분석결과, 최종적으로 3개의 변수가 선택되었으며(4) 변수의 추가에 따른 모형의 추가 설명력 F검정은 모두 유의했다. 각각의 예측계수 밑에 기재된 값은 추정량의 표준편차이다. putting_avg가 1 작아지면 rank가 평균적으로 상승하고, green_in_regulation과 birdies의 경우 1만큼 증가하면 rank를 평균적으로 각각 올려준다.

	Dependent variable:							
	rank					prize		
변수선택 단계	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
birdies								
green_in_regulation								
putting_avg								
Constant								
Observations	85	85	85	85	85	85	85	85
R2								
Adjusted R2								
Residual Std. Error								
F Statistic								
Note:	*p<0.1; **p<0.05; ***p<0.01							

2) 모형의 진단

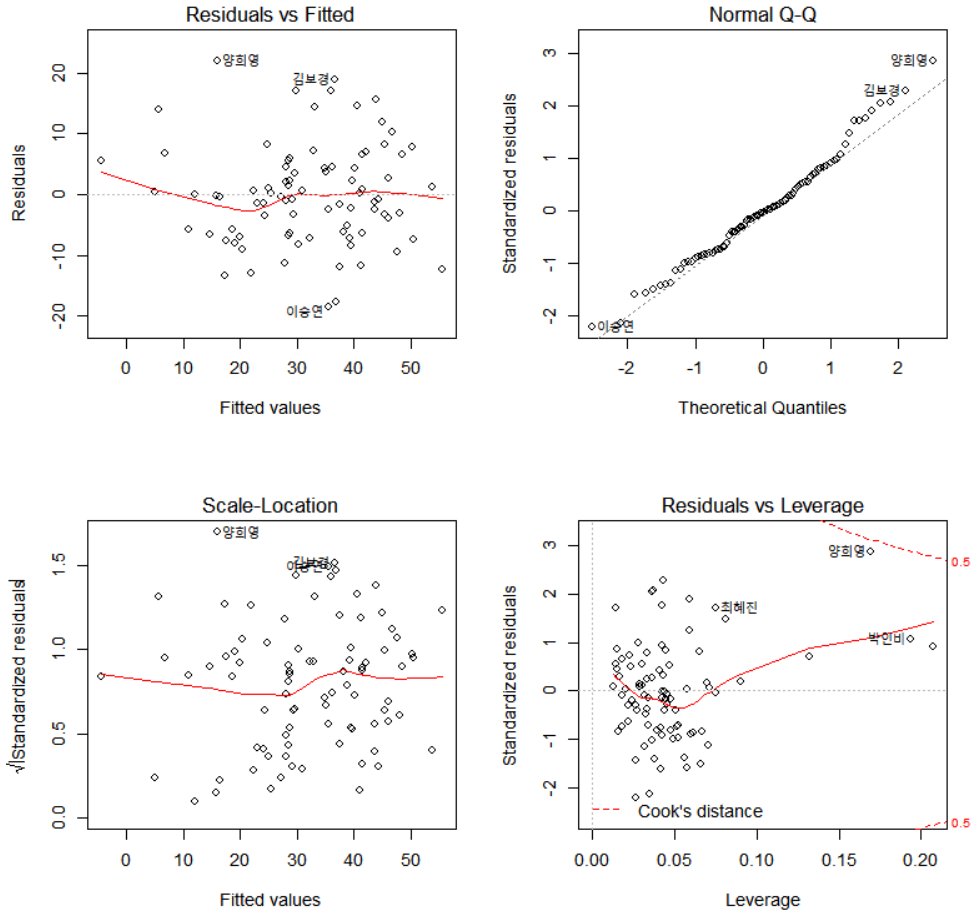
다음으로 선형회귀모형의 이론적 가정을 검증하기 위해 모형의 진단을 진행한다. 여기서는 아래 4개의 그래프가 그려진 그림에서 각 그래프가 의미하는 바를 설명한다.

가) 선형성 - 왼편 위: 잔차와 예측된 값 사이에 함수관계가 없어야 한다. 즉, 모형은 무작위 오차를 제외하고 데이터에 존재하는 모든 체계적 변량을 포착해야 한다. 그림을 보면 예측값과 잔차 사이의 명백한 함수관계가 보이지 않으므로 선형성을 위배하지 않는다.

나) 잔차의 정규성(Normality) - 오른쪽 위: 모형의 잔차가 정규분포를 따르는지 정규 QQ 플랏을 통해 확인한다. 점들이 일직선상에 놓여있으면서 크게 벗어나지 않으므로 정규성을 위배하지 않는다.

다) 등분산성 – 왼편 아래: 잔차가 등분산성을 만족한다면 Scale-Location 그래프에서 점들이 수평선 주변에 무작위로 분포해야 한다. 그림을 통해 판단해볼 때 등분산성 또한 만족한다.

라) 이상치(outlier), 영향관측치(influential) - 오른편 아래: 모형이 설명하는 관측치 중에 극단적인 관측치가 없어야 하고, Cook의 거리에 따라 계수의 추정에 영향을 주는 영향관측치가 없어야 한다. 이를 기준으로 이상치로 판단되는 관측값은 양희영, 김보경, 이승연 세 건이 있다. 김보경, 이승연 선수의 관측치는 이상값에 해당하지만 모형의 적합에 영향을 유의하게 주지 않으므로 제거하지 않는다. 반면 양희영 선수의 경우 Cook's distance를 볼 때 임계점인 0.5에 근접하므로 유의미한 영향관측치이므로 제거하고 모형을 재적합한다.



마) 잔차의 독립성:

lag	Autocorrelation	D-W Statistic	p-value
1			

위 검정결과에서 잔차의 Durbin-Watson 자기상관성 통계량의 p-value가 0.05보다 크므로 관측치들이 자기상관성이 없다. 잔차의 독립성 가정을 만족한다.

바) 분산팽창지수(VIF):

birdies	green_in_regulation	putting_avg

선형회귀모형은 독립변수 간 다중공선성이 존재할 때 추정량의 신뢰성이 부족하다. 서로 강한 선형관계가 있는 변수들의 회귀계수 해석은 난해하기 때문이다. 특히 VIF 값이 5를 넘으면 다중공선성이 문제가 된다. 이 경우엔 다중공선성에 따른 모형의 신뢰성에 문제가 없는 것으로 판단한다.

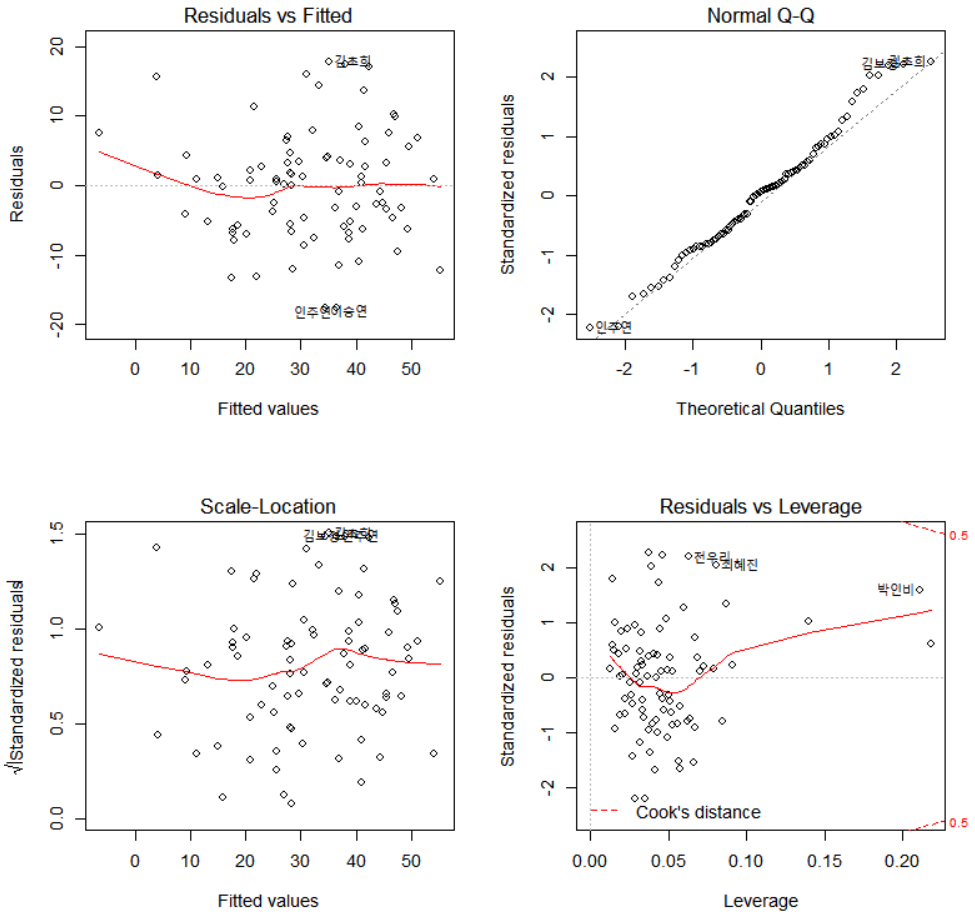
3) **모형의 진단결과에 따른 재적합 및 최종모형**

모형의 재적합 결과 변수들의 효과크기가 조정되었다. putting_avg의 효과가 조정되었고, reen_in_regulation과 birdies의 경우 각각 으로 조정되었다. 이는 양희영 선수의 경기력이 birdies, green_in_regulation, putting_avg을 기준으로 대체적으로 평균보다 좋았지만 rank는 낮아서 계수의 추정에 영향을 주었기 때문으로 보인다.

	Dependent variable:			
	rank			
변수선택 단계	(1)	(2)	(3)	(4)
birdies				
green_in_regulation				
putting_avg				
Constant				
Observations	84	84	84	84
R2				
Adjusted R2				
Residual Std. Error				
F Statistic				
Note:	*p<0.1; **p<0.05; ***p<0.01			

4) **재적합 모형의 진단**

재적합 결과 모형의 진단 결과는 전술한 OLS 회귀분석의 가정을 모두 만족하는 것으로 판단함.



나. 주제 3. LPGA 시즌결과물(rank)의 (최대)결정요인 탐색

1) 전진선택

rank의 분석결과, 최종적으로 3개의 변수가 선택되었으며(4) 변수의 추가에 따른 모형의 추가 설명력 F검정은 모두 유의했다. green_in_regulation이 1 증가하면 rank가 평균적으로 상승하고, putting_avg가 1 감소하면 rank가 평균적으로 상승한다. driving_distance의 경우 1만큼 증가하면 rank가 평균적으로 상승한다.

	Dependent variable:							
	rank						prize	
변수선택 단계	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
green_in_regulation								
putting_avg								
driving_distance								
Constant								
Observations	86	86	86	86	86	86	86	86
R2								
Adjusted R2								
Residual Std. Error								
F Statistic								
Note:	*p<0.1; **p<0.05; ***p<0.01							

2) **모형의 진단**

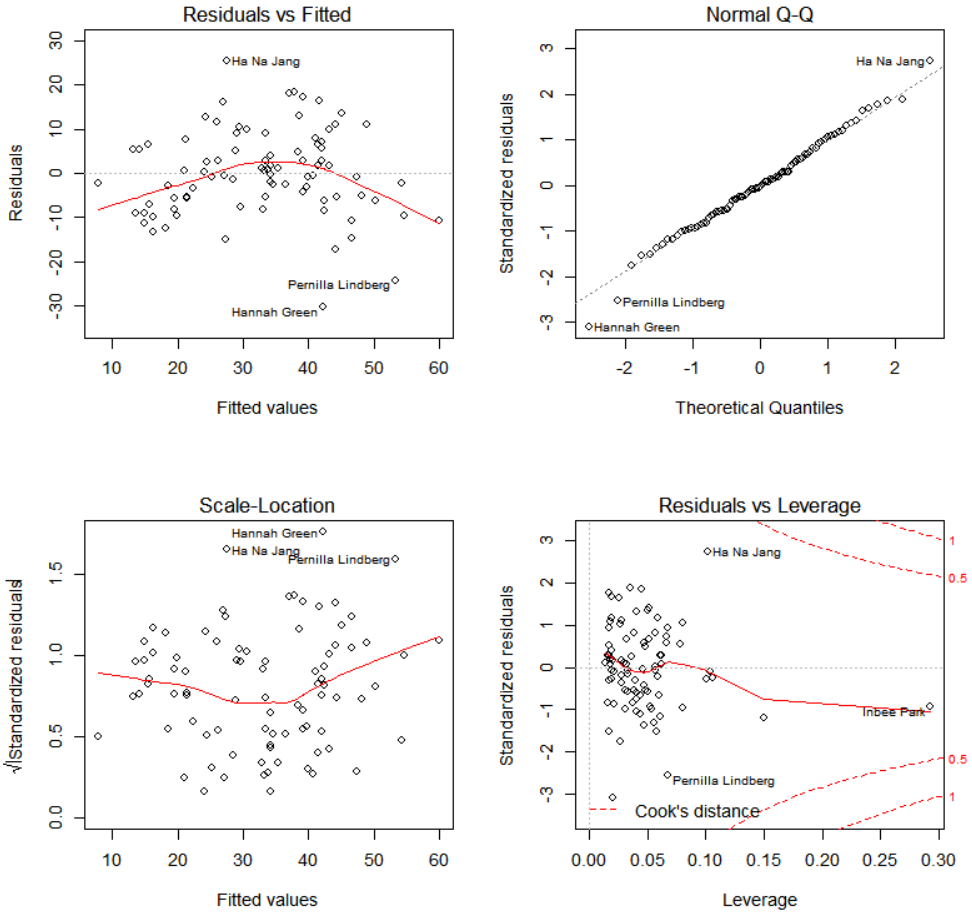
마찬가지로 선형회귀모형의 이론적 가정을 검증하기 위해 모형의 진단을 진행한다.

가) 선형성 - 왼편 위: 그림을 보면 예측값과 잔차 사이의 명백한 함수관계가 보이지 않으므로 선형성을 위배하지 않는다.

나) 잔차의 정규성(Normality) - 오른쪽 위: 모형의 잔차가 정규분포를 따르는지 정규 QQ 플랏을 통해 확인한다. 점들이 일직선상에 놓여있으면서 크게 벗어나지 않으므로 정규성을 위배하지 않는다.

다) 등분산성 – 왼편 아래: 그림을 통해 판단해볼 때 등분산성 또한 만족한다.

라) 이상치(outlier), 영향관측치(influential) - 오른편 아래: 이상치로 판단되는 관측값은 Ha Na Jang, Pernilla Lindberg, Hannah Green이다. 이들 관측치는 이상값에 해당하지만 모형의 적합에 영향을 유의하게 주지 않으므로 제거하지 않는다.



마) 잔차의 독립성:

lag	Autocorrelation	D-W Statistic	p-value
1			

잔차의 Durbin-Watson 자기상관성 통계량의 p-value가 0.05보다 크므로 관측치들이 자기상관성이 없다. 잔차의 독립성 가정을 만족한다.

바) 분산팽창지수(VIF):

driving_distance	green_in_regulation	putting_avg

다중공선성에 따른 모형의 신뢰성에 문제가 없는 것으로 판단한다. 따라서 최종모형을 (4)로 확정한다.

다. 주제4. KLPGA + LPGA 시즌결과물의 (최대)결정요인 탐색

1) 전진선택

rank의 분석결과, 최종적으로 3개의 변수가 선택되었으며(4) 변수의 추가에 따른 모형의 추가 설명력 F검정은 모두 유의했다. green_in_regulation이 1 증가하면 rank가 평균적으로 상승하고, putting_avg가 1 감소하면 rank가 평균적으로 상승한다. birdies의 경우 1만큼 증가하면 rank가 평균적으로 상승한다.

	Dependent variable:									
	rank				prize					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
green_in_regulation										
putting_avg										
birdies										
driving_accuracy										
driving_distance										
Constant	32.50***	220.70***	-58.37	-110.69***	5.98***	-62.57***	42.79**	70.45***	65.50***	49.30***
	(1.13)	(21.91)	(38.39)	(39.89)	(0.37)	(7.10)	(17.13)	(15.09)	(14.81)	(18.42)
Observations	171	171	171	171	171	171	171	171	171	171
R2										
Adjusted R2										
Residual Std. Error										
F Statistic										
Note:	*p<0.1; **p<0.05; ***p<0.01									

2) **모형의 진단**

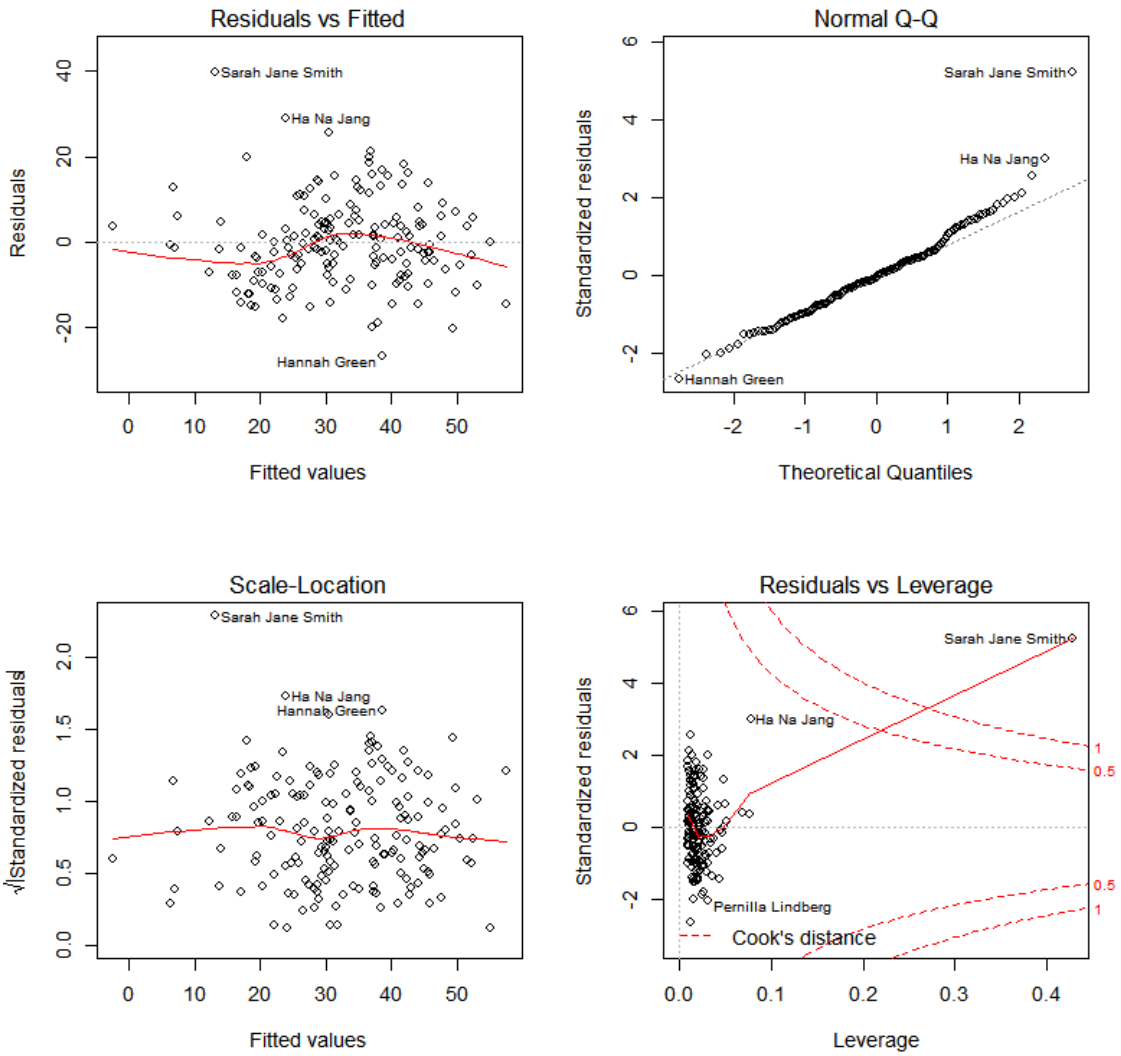
마찬가지로 선형회귀모형의 이론적 가정을 검증하기 위해 모형의 진단을 진행한다.

가) 선형성 - 왼편 위: 그림을 보면 예측값과 잔차 사이의 명백한 함수관계가 보이지 않으므로 선형성을 위배하지 않는다.

나) 잔차의 정규성(Normality) - 오른쪽 위: 모형의 잔차가 정규분포를 따르는지 정규 QQ 플랏을 통해 확인한다. 점들이 일직선상에 놓여있으면서 크게 벗어나지 않으므로 정규성을 위배하지 않는다.

다) 등분산성 – 왼편 아래: 그림을 통해 판단해볼 때 등분산성 또한 만족한다.

라) 이상치(outlier), 영향관측치(influential) - 오른편 아래: 이상치로 판단되는 관측값은 Ha Na Jang, Sarah Jane Smith, Hannah Green이다. 강력한 영향관측치는 Sarah Jane Smith이고, 이 관측치를 제거하고 재적합하기로 한다.



마) 잔차의 독립성:

lag	Autocorrelation	D-W Statistic	p-value

잔차의 Durbin-Watson 자기상관성 통계량의 p-value가 0.05보다 크므로 관측치들이 자기상관성이 없다. 잔차의 독립성 가정을 만족한다.

바) 분산팽창지수(VIF):

green_in_regulation	putting_avg	birdies

다중공선성에 따른 모형의 신뢰성에 문제가 없는 것으로 판단한다. 따라서 최종모형을 (4)로 확정한다.

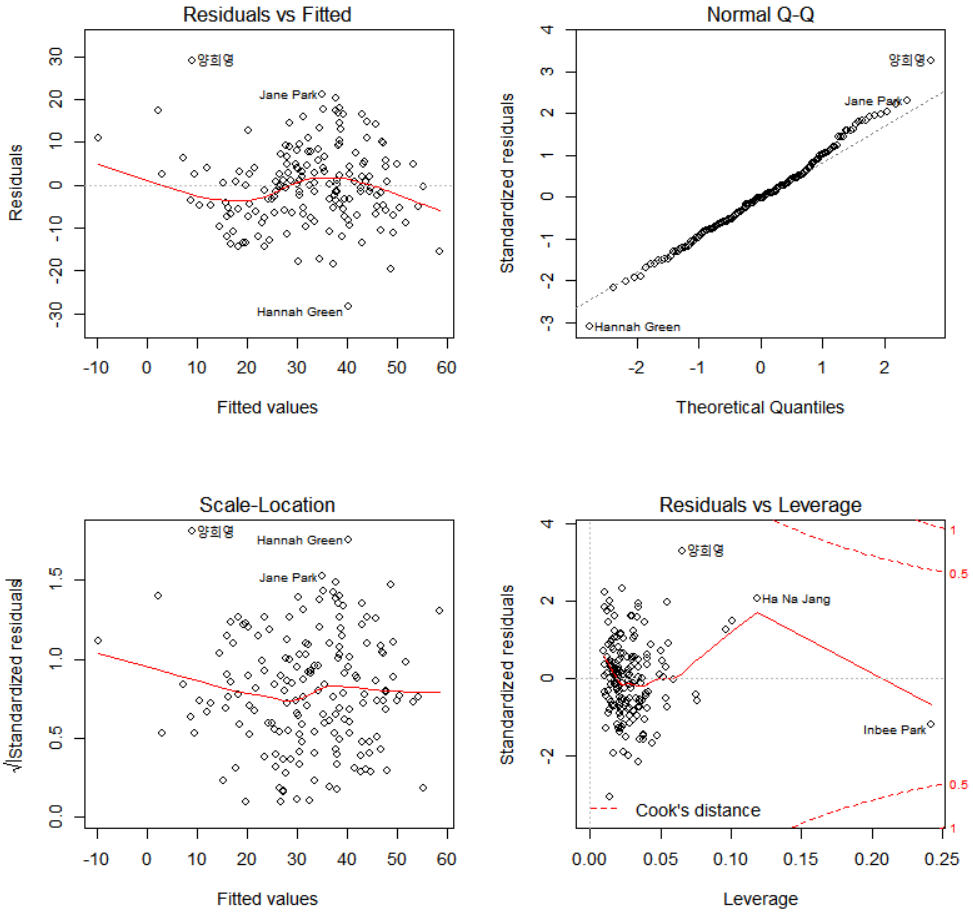
3) **모형의 진단결과에 따른 재적합 및 최종모형**

모형의 재적합 결과 변수들의 효과크기가 조정되었다. putting_avg의 효과가 조정되었고, green_in_regulation과 birdies의 경우 각각 birdies는 조정되었다. 마지막으로 driving_distance가 1 증가하면 상승하는 효과가 추가되었다. 이는 선수의 경기력이 평균보다 굉장히 우수했던 반면 rank가 으로 낮은 것에서 기인한 모형의 편의(bias)였던 것으로 분석된다.

	Dependent variable:				
	rank				
	(1)	(2)	(3)	(4)	(5)
green_in_regulation					
putting_avg					
birdies					
driving_distance					
Constant	32.38*** (1.13)	219.57*** (21.87)	-59.20 (38.23)	-179.67*** (38.57)	-127.83*** (46.40)
Observations					
R2					
Adjusted R2					
Residual Std. Error					
F Statistic					
Note:	*p<0.1; **p<0.05; ***p<0.01				

4) **재적합 모형의 진단**

재적합 결과 모형의 진단 결과는 전술한 OLS 회귀분석의 가정을 모두 만족하는 것으로 판단함. 여기서도 선수가 이상치 및 영향관측치 후보로 나타났지만, KLPGA 분석과는 다르게 Cook's distance 기준으로 임계값 0.5에 근접하지 않으므로 제거하지 않고 모형(5)를 최종모형으로 확정한다.



4. 전진선택법(forward-selection)을 이용한 로지스틱 회귀분석

가. 주제 5, 6. 우승경력 有/無 선수들의 시즌결과물(KLPGA, LPGA) (최대)결정요인 탐색

1) **로지스틱 회귀분석**

로지스틱 회귀분석은 반응변수인 y가 독립변수가 주어졌을 때, 서로 독립인 베르누이 분포를 따른다고 가정한다. 이때 베르누이 분포의 평균인 p, 즉 관심사건의 발생확률(우승경력이 생길 확률)을 선형모형의 함수로 추정하는 모형이 로지스틱 회귀분석이다. 이를 수식으로 나타내면, $\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right)=\beta_0+\mathrm{x}_1\beta_1+\ldots+\mathrm{x}_\mathrm{p}\beta_\mathrm{p}$로 표현된다.

로지스틱 회귀분석의 회귀계수의 해석은 다음과 같다. $e^{\beta}=\frac{p_1/(1-p_1)}{p_0/(1-p_0)}$

위 수식에서 p_1, p_0는 각각 독립변수가 1, 0일 때 종속변수의 발생 확률이다. 즉, 독립변수가 한 단위 증가할 때 승수(Odds)의 비(Ratio)로 해석되는 것이다. 예를 들어 추정 회귀계수가 1이라면 $e^1 \approx 2.72$ 이 승수비 값이 되고, 이때 해석은 독립변수가 한 단위 증가할 때 관심사건이 발생할 승수가 2.72배 커진다는 의미이다.

본 분석에선 종속변수를 우승경력(1 = 있음, 0 = 없음)으로 재코딩하여 분석했다.

2) 모형의 적합

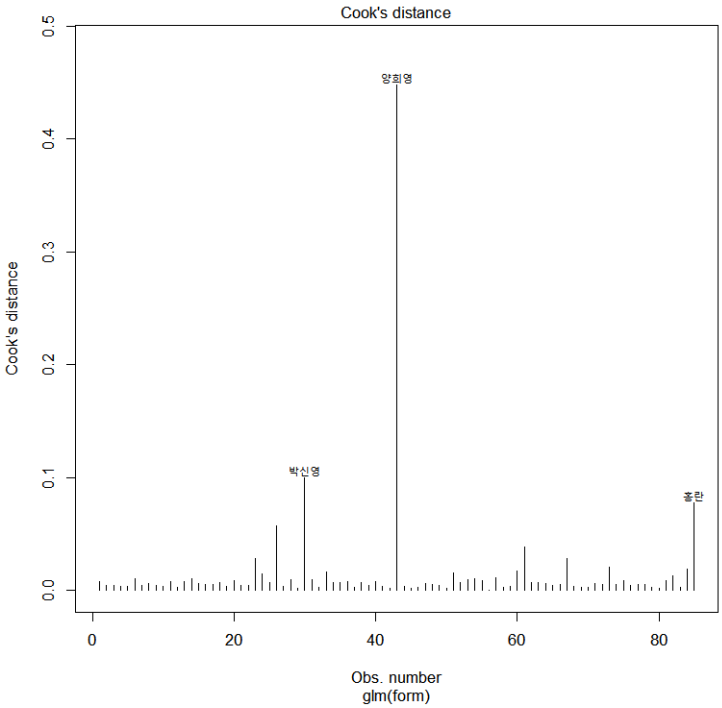
모형 적합결과 KLPGA는 birdies 하나, LPGA는 birdies, green_in_regulation, putting_avg, driving_accuracy가 최종모형의 변수들로 선택되었다. 선형회귀분석에선 추가 제곱합이 유의하게 증가하면 변수를 선택하지만, 일반화선형모형에서는 변수를 추가할 때 Akaike Information Crierion(AIC) 값을 감소시키면 모형의 적합도가 증가한다고 보고 변수를 선택한다. KLPGA는 birdies가 1 증가하면 선수가 우승할 승수가 $e^{0.66} \approx 1.93$으로 커졌다. 반면 birdies가 1 작아지면 우승할 승수는 감소한다. LPGA는 birdies가 통계적으로 유의미한 변수는 아니지만(p-value > 0.1), 1 증가할 경우 반대로 우승할 승수가 줄어들었다. 반면 1 작아질 경우 43% 증가한다. green_in_regulation은 1만큼 커질때 우승할 승수를 증가시키고, putting_avg, driving_accuracy는 1만큼 커질경우 각각 감소시켰다. 따라서 green_in_regulation은 값이 클수록 우승할 확률을 높여주는 기술력이며 나머지는 값이 클수록 우승을 못할 확률을 높여주는데 특히 putting_avg의 경우 우승할 승수를 극단적으로 감소시킨다. 반대로 putting_avg, driving_accuracy가 1 줄어들 경우 우승할 승수는 각각 증가시킨다.

특이점은 KLPGA에선 오직 birdies만 선수의 우승 여부에 유의미한 영향력을 가지는 변수인 반면, LPGA에선 birdies를 제외한 모든 경기력 변수가 통계적으로 유의미한 영향을 가진 것으로 분석됐다. 선형회귀분석과 달리 로지스틱 회귀분석은 변수의 선택에 따른 모형의 적합도 증가를 결정계수가 아닌 Akaike Information Criterion(AIC)을 본다. AIC값이 작아지면 모형의 적합도가 개선되었다고 판단하므로, 두 모형 모두 최적의 모형을 탐색한 결과이다.

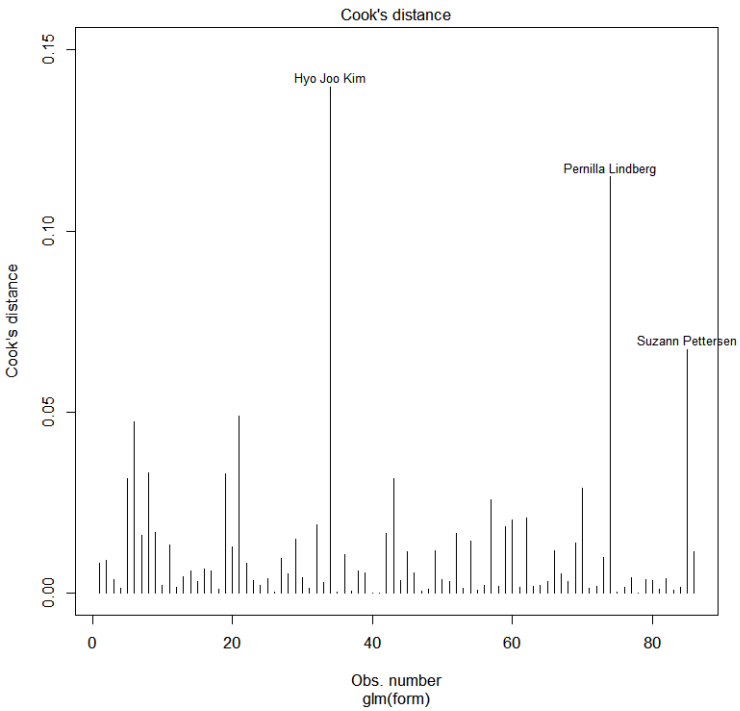
	Dependent variable: wins						
	KLPGA		**LPGA**				
	(1)	(2)	(1)	(2)	(3)	(4)	(5)
birdies							
green_in_regulation							
putting_avg							
driving_accuracy							
Constant							
Observations	85	85	86	86	86	86	86
Log Likelihood							
Akaike Inf. Crit.							
Note:	*p<0.1; **p<0.05; ***p<0.01						

3) **모형 진단**

선형회귀분석과 달리 로지스틱 회귀분석의 경우 잔차에 대한 이론적 가정을 하지 않기 때문에 영향관측치만 진단 대상으로 정한다. 영향관측치는 그림에서 Cook's distance가 0.5보다 크거나 가까운 값을 기준으로 한다. KLPGA의 양희영 선수의 관측치가 영향관측치로 판정됐다. 따라서 이 관측치를 제거하고 모형을 재적합 하기로 한다.



LPGA의 경우 유의미한 이상치 및 영향관측치는 관찰되지 않았다. 따라서 최종모형은 (5)로 선정한다.



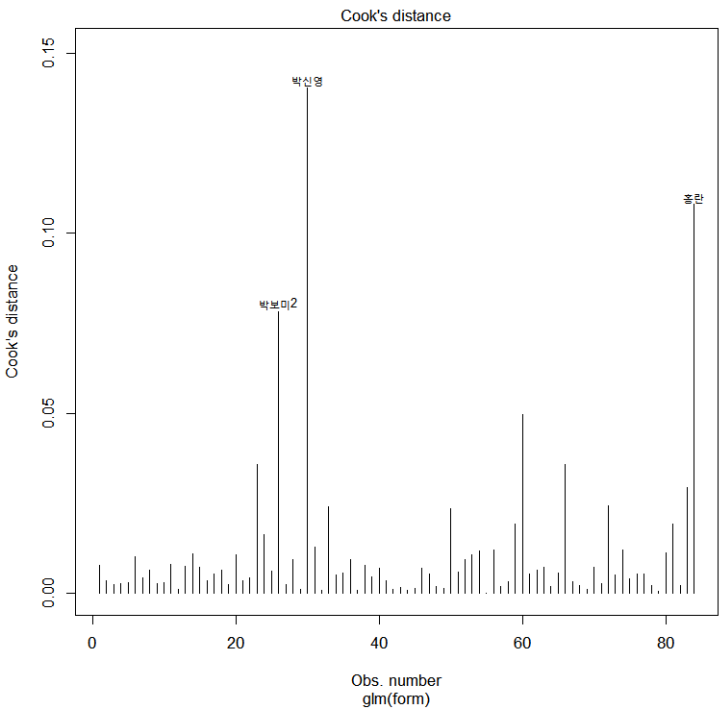
4) 모형의 진단결과에 따른 재적합 및 최종모형

분석결과 birdies의 효과가 우승 승수의 증가율 기준으로 93%에서 $e^{0.84} \approx 2.32$이므로 조정되었다. 반면 birdies가 한 단위 감소할 경우 우승의 승수는 감소한다.

	Dependent variable: wins	
	KLPGA	
	(1)	(2)
birdies		
Constant		
Observations	84	84
Log Likelihood		
Akaike Inf. Crit.		
Note:	*p<0.1; **p<0.05; ***p<0.01	

5) 재적합 모형의 진단

재적합 결과 유의미한 영향관측치는 관찰되지 않았다. 따라서 위 모형을 최종모형으로 채택한다.



5. 앙상블(Ensemble) 기법을 이용한 로지스틱 회귀분석 예측모형 개발

가. Boosting 알고리즘

부스팅 알고리즘은 성능이 낮은 모형(weak learner, 약한 학습기)을 반복적인 알고리즘을 통해 예측력을 강화시켜주는 머신러닝 알고리즘이다. 가장 대표적인 부스팅 알고리즘은 Adaboost(Adaptive Boosting)이며, 최근 많은 연구를 통해 Gradient Boost, XGboost 등의 형태로 다양하게 응용되고 있다. 여기서는 부스팅 알고리즘은 지난 수십년 간 학계의 이론 및 실증연구를 통해 전통적인 분류 혹은 회귀모형, 의사결정나무 등의 약한 학습기의 예측성능을 비약적으로 끌어올려주는 것으로 검증되었다. 이에 Friedman et al.(2000), Friedman(2001)의 연구를 참고해 XGBoost를 이용한 로지스틱 예측모형을 개발하고자 한다. XGBoost의 구현에는 R 패키지 xgboost를 사용하였다.

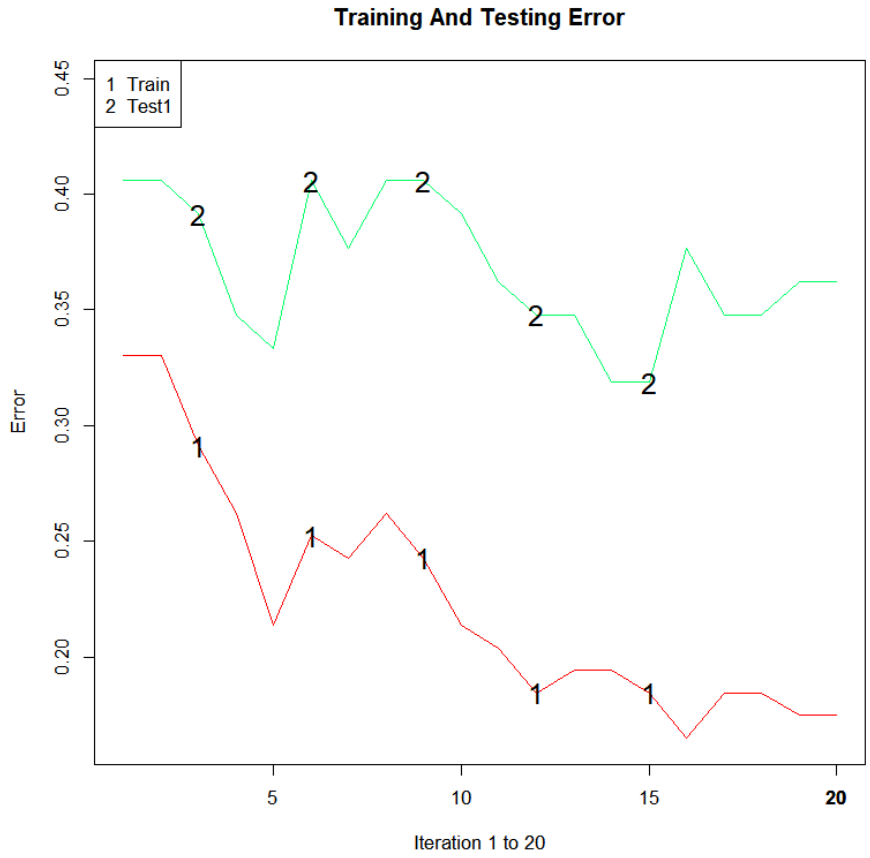
나. 우승경력 無 top10 ranker 중 향후 우승 가능성 예측

1) 예측대상 선수들

예측대상 선수들은 상위 10위 랭커들 중 우승경력이 없는 선수들이다. 예측용 데이터를 생성하기 위해 먼저 루키로 선정된 해를 기준으로 career 변수를 생성하였다. 2020년 상반기 기준으로 수집된 데이터이기 때문에 그 값에 0.5를 더해줬다. career 변수를 추가해 XGBoost 알고리즘으로 우승선수를 예측했다.

2) performance plot

부스팅 결과 최종모형의 test data에 대한 성능은 오분류율(error rate) %로 나타났다.



참고문헌

[1] Friedman, J., Hastie, T., &Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

[2] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.