2018 Fall

Computational Statistics
HW#2

182STG18 이하경

[Problem] Variable Selection in Regression

Goal Stepwise Regression 과 All possible Regression 을 사용해 회귀 모형에서 설명변수 선택에 따른 AIC 값을 비교하고 가장 작은 AIC 값을 가지는 최적 모형을 찾는다.

Baseball Data

Response		
log(salary)		
Potential Predictors		
average	obp	runs
hits	doubles	triples
homeruns	rbis	walks
sos	sbs	errors
freeagent	arbitration	runsperso
hitsperso	hrsperso	rbisperso
walksperso	obppererror	runspererror
hitspererror	hrspererror	soserrors
sbsobp	sbsruns	sbshits

데이터는 총 337 명의 야구선수의 연봉과, 경기 실적에 관련된 27 가지 속성을 변수로 가지고 있다. 선수들의 연봉의 로그 값을 종속변수로, 27 가지 잠재적인 설명변수로 하여 최적 선형 회귀 모형을 찾아보려고 한다.

예측력을 가장 좋게 하기 위한 설명변수의 조합, optimal p*를 찾기 위해 변수 선택 방법으로 AIC 값을 기준으로 한 Stepwise Regression 과 R 의 leaps package 를 이용한 All Possible Regression 을 각각 시행하고 두 가지 방법의 결과를 비교한다.

모형의 p 가 증가할수록 RSS 는 계속해서 감소하나 AIC 의 경우 모형의 복잡성에 대한 Penalty 가 더해지므로 과대적합을 방지할 수 있다. AIC 가 가장 작은 모형을 선택할 경우 Test Data 에서의 예측력이 가장 좋을 것으로 예상한다.

Method 1. Stepwise Regression

다음 왼쪽의 표는 Stepwise Selection 을 시행하기 전 먼저 27 개의 가능한 설명변수를 모두 사용하여 적합한 Full Model 의 결과이다. 각 변수의 p-value 값을 보면 다른 변수들이 모형에 존재할 때 유의수준 0.05 에서 추가 설명력을 가지는 변수는 5 개로 매우 적은 편이다. 따라서 불필요한 설명변수를 제거하여 모형의 복잡성을 줄이는 것이 효율적일 것이라고 예상한다. 오른쪽 표에서 Full Model 의 AIC 값은 -396.71 이다.

Summary of Full Model							
RSS=87.948 (df=309), Adjusted R^2=79.44%							
	est.	**se**	**p-value**		**est.**	**se**	**p-value**
(**Intercept**)	5.381	0.275	**< 0.001**	**arbitration**	1.348	0.089	**< 0.001**
average	-1.073	2.669	0.688	runsperso	-0.317	0.206	0.123
obs	-0.391	2.381	0.870	**hitsperso**	0.312	0.134	**0.021**
runs	0.016	0.007	**0.017**	hrsperso	1.035	0.615	0.093
hits	-0.005	0.004	0.197	rbisperso	-0.430	0.265	0.105
doubles	0.003	0.007	0.667	walksperso	-0.183	0.235	0.437
triples	-0.015	0.018	0.395	obppererror	-0.724	0.768	0.347
homeruns	-0.014	0.015	0.350	runspererror	-0.005	0.013	0.688
rbis	0.018	0.006	**0.005**	hitspererror	0.007	0.008	0.386
walks	0.003	0.004	0.428	hrspererror	-0.006	0.02	0.768
sos	-0.006	0.003	0.064	soserrors	0.000	0.000	0.656
sbs	-0.008	0.031	0.796	sbsobp	0.094	0.099	0.344
errors	-0.002	0.016	0.884	sbsruns	0.000	0.000	0.230
freeagent	1.509	0.077	**< 0.001**	sbshits	0.000	0.000	0.912

AIC of Full/Null Model		
Model	**df(=p)**	**AIC**
Full	28	-396.71
Null	1	110.58

cf. $\mathbf{AIC} = \mathbf{n}\log\frac{\mathbf{RSS}}{\mathbf{n}} + \mathbf{2p}$

for Linear Reression Model

p	Variable In/Out	AIC
28		-396.71
27	- sbshits	-398.69
26	- errors	-400.67
25	- obp	-402.63
24	- sbs	-404.57
23	- hrspererror	-406.42
22	- doubles	-408.23
21	- runspererror	-409.83
20	- walks	-411.25
19	- walksperso	-412.80
18	- homeruns	-413.58
17	- hits	-414.83
16	- rbisperso	-415.82
15	- hrsperso	-416.49
14	- hitspererror	-417.26
13	- obppererror	**-418.94**
12	- triples	-418.45

Summary of Stepwise Model

RSS=89.997 (df=324), Adjusted R^2=79.93%

	est.	se	p-value		est.	se	p-value
(**Intercept**)	5.290	0.228	**< 0.001**				
average	-1.701	0.976	0.082	**arbitration**	1.329	0.085	**< 0.001**
runs	0.015	0.003	**< 0.001**	**runsperso**	-0.391	0.092	**< 0.001**
triples	-0.023	0.015	0.122	**hitsperso**	0.171	0.051	**0.001**
rbis	0.011	0.002	**< 0.001**	**soserrors**	0.000	0.000	**0.027**
sos	-0.005	0.002	**0.022**	**sbsobp**	0.075	0.030	**0.013**
freeagent	1.494	0.073	**< 0.001**	**sbsruns**	0.000	0.000	**0.033**

Null Model 부터 Full Model 까지의 범위에서 AIC 기준의 Stepwise Regression 을 시행한 결과 각 단계에서 제외(또는 선택)된 변수와 AIC 값의 변화는 다음과 같다. Intercept 를 포함한 p=28 의 Full Model 을 시작으로 p=13 수준까지 각 단계에서 변수가 하나씩 제외되었다. 다음 단계에서 변수를 제외(또는 추가)할 경우 AIC 가 다시 증가하므로, min AIC 값을 가질 것으로 기대되는 최종 모형은 12 개의 설명변수를 가진 p=13 모형이며 이때의 AIC 는 -418.94 이다. 모형의 통계량을 확인해보면 불필요한 설명변수들이 대부분 제거되고 회귀계수들의 p-value 값이 대부분 유의수준 0.05 보다 작다. 따라서 Full Model 에 비해 모형의 복잡성은 감소하면서도 예측력이 좋을 것으로 예상되는 모형이라고 할 수 있다.

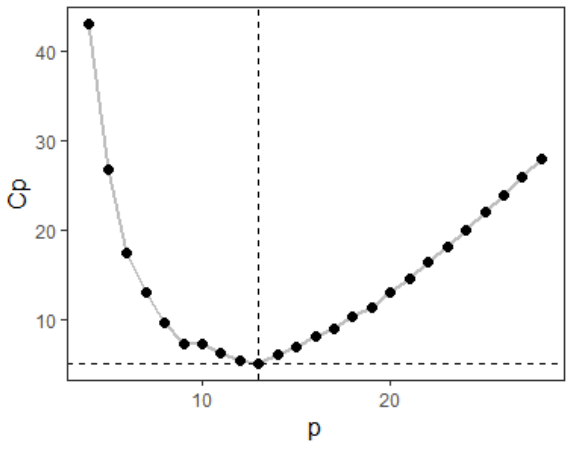
Method 2. All Possible Regression

가능한 모든 설명변수의 조합을 고려하기 위해 R 의 leaps package: regsubsets 함수를 이용해 설명변수의 개수가 1 부터 27 일 때 각 p 수준에서 RSS 를 가장 작게 하는 변수의 조합을 찾고, 각각의 모형에서의 Mallow's Cp* 값을 확인하였다.

$$* \mathrm{Cp} = \frac{\mathrm{RSS}}{\sigma^2} - \mathrm{n} + 2\mathrm{p}$$

Cp 가 가장 작은 모형이 AIC 가 가장 작은 모형과 같을 것으로 예상하고, 해당 기준으로 하위 3 개의 모형을 선택하고 AIC 값을 확인하였다.

아래의 표에서 각각의 p 수준에서의 선택된 변수들과 Cp 값을 확인할 수 있으며, 오른쪽 그래프는 Cp 값이 매우 큰 p=2, 3 을 제외하고 나머지 p 수준에 따른 Cp 값의 변화를 나타낸 것이다.



p	Selected Variables																									Cp
2				●																						572.98
3												●	●													244.85
4							●					●	●													43.10
5				●			●					●	●													26.82
6				●			●					●	●									●				17.51
7			●				●		●			●	●	●												13.04
8			●				●		●			●	●	●	●											9.81
9			●				●		●			●	●					●	●	●						7.34
10			●				●		●			●	●					●	●	●		●				7.29
11			●				●		●			●	●	●	●							●	●	●		6.37
12	●		●				●		●			●	●	●	●							●	●	●		5.55
13		●	●			●		●		●			●	●	●	●							●	●	●		**5.20**
14	●		●			●		●		●			●	●	●	●	●						●	●	●		6.12
15		●	●			●		●	●	●			●	●	●	●	●						●	●	●		7.09
16		●	●			●		●	●	●			●	●	●	●	●	●					●	●	●		8.13
17		●	●	●		●	●	●	●	●			●	●	●	●	●	●					●	●	●		9.00
18		●	●	●		●	●	●	●	●			●	●	●	●	●	●					●	●	●		10.48
19		●	●	●		●	●	●	●	●		●	●	●	●	●	●	●		●		●			●	●		11.42
20		●	●	●		●	●	●	●	●		●	●	●	●	●	●	●		●		●			●	●		13.08
21	●		●	●		●	●	●	●	●		●	●	●	●	●	●	●	●	●		●			●	●		14.65
22	●		●	●		●	●	●	●	●			●	●	●	●	●	●	●	●	●	●		●	●	●		16.44
23	●		●	●	●	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●		●	●	●		18.26
24	●		●	●	●	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●	●	●		20.13
25	●		●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●		22.07
26	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●		24.03
27	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		26.01
28	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	28.00

총 27 개의 모형 중 Cp 값이 5.20 으로 가장 작은 모형은 p=13, 설명변수가 총 12 개인 모형이다. Cp 기준의 하위 3 개 모형의 설명변수들로 적합한 회귀 모형의 AIC 값을 확인해본 결과 AIC 역시 p=13 수준에서 가장 작았다. 따라서 All Possible Regression 방법으로 선택된 최종 모형은 2, 3, 6, 8, 10, 13, 14, 15, 16, 24, 25, 26 번 설명변수로 하는 모형이다.

p	Cp	AIC
12	5.55	-418.45
13	**5.20**	**-418.95**
14	6.12	-418.09

Summary of All Possible Model (p=13)

RSS=89.997 (df=324), Adjusted R^2=79.93%

	est.	se	p-value		est.	se	p-value
(Intercept)	5.293	0.230	**< 0.001**				
obp	-1.345	0.772	0.082	arbitration	1.327	0.085	**< 0.001**
runs	0.016	0.003	**< 0.001**	runsperso	-0.369	0.091	**< 0.001**
triples	-0.026	0.015	0.078	hitsperso	0.150	0.048	**0.002**
rbis	0.010	0.002	**< 0.001**	soserrors	0.000	0.000	**0.021**
sos	-0.004	0.002	**0.030**	sbsobp	0.077	0.030	**0.010**
freeagent	1.508	0.072	**< 0.001**	sbsruns	0.000	0.000	**0.026**

[Discussion]

- Summary -

Method	Selected Variables																		p	AIC
	1	2	3	4	6	8	9	10	13	14	15	16	20	21	22	24	25	26		
All Possible		•	•		•	•		•	•	•	•	•				•	•	•	13	**-418.95**
Stepwise	•		•		•	•		•	•	•	•	•				•	•	•	13	-418.94
(+) Forward			•	•		•	•	•	•	•	•		•	•	•	•	•	•	15	-413.02

Baseball Data 에서 min AIC 를 찾기 위해 고려 가능한 총 모형의 개수는 2^{27}=134,217,728 개로 매우 많다. All Possible Regression 은 각각의 p 수준 내에서는 RSS 를 가장 작게 하는 변수 조합을 선택하고 이들 27 개 모형의 Cp 값을 비교하는 방법으로 가능한 모든 경우를 고려하지만, Stepwise Regression 은 각 단계의 이전과 다음 step 에서의 AIC 값을 비교하여 빠른 시간 내에 local optimum 값을 찾아내는 'Local Search' 방법이므로 본 과제의 경우처럼 global minimum 값을 찾아내지는 못하는 경우가 발생하기도 한다. 두 가지 방법에서 선택된 모형의 p 는 Intercept 를 포함해 13 으로 동일하지만 선택된 설명변수가 완전히 같지는 않았으며, All Possible 방법에서 선택된 모형의 AIC 가 0.01 의 차이로 미세하게 더 작았다. 값의 차이가 크지 않으므로 가능한 모든 경우의 수를 다 고려하기보다 빠른 시간 내에 합리적인 값을 찾아내는 것이 더 효율적일 수 있다.

Stepwise 방법은 각 단계에서 이전에 제거된 설명변수의 재선택을 고려하나 이 경우는 재선택되지 않았으므로 Backward Elimination 을 이용했을 때와 결과적으로 같았다. 이와 다르게 Forward Selection 을 이용한 변수 선택을 시행하여 보았을 때 최종 선택된 변수의 개수와 조합은 앞의 두 가지 방법과 상이했으며 AIC 값은 조금 더 컸다. 모든 방법에서 공통적으로 선택된 설명변수는 3, 8, 10, 13, 14, 15, 24, 25, 26 번 변수로 해당 변수들이 야구선수의 연봉을 설명하는 중요한 변수임을 알 수 있다.

[Appendix] R Code

```
library(dplyr)
library(leaps)
library(ggplot2)

  baseball <- read.table("D:₩₩2 학기₩₩통계계산특론 1₩₩HW2₩₩baseball.txt", header = T, sep = " ")
  str(baseball)
  summary(baseball)

# step
  full <- lm(log(salary) ~ ., baseball) ; summary(full) ; anova(full)
  null <- lm(log(salary) ~ 1, baseball) ; summary(null)

  extractAIC(full)
  extractAIC(null)

  bothstep <- step(full, direction = "both") ; summary(bothstep) ; anova(bothstep)
  backstep <- step(full, direction = "backward") ; summary(backstep)
  forstep <- step(null, scope = list(lower=null, upper=full), direction = "forward") ; summary(forstep) ; anova(forstep)

  extractAIC(bothstep)
  extractAIC(backstep)
  extractAIC(forstep)


# all possible
  allsubs <- regsubsets(log(salary) ~ ., data=baseball, nvmax=27)
  summary(allsubs)

  plot(allsubs, scale = "Cp")

  allsubsinfo <- data.frame(p=2:28, summary(allsubs)$outmat, Cp=summary(allsubs)$cp)
  allsubsinfo %>% filter(rank(Cp) %in% 1:3)

  sub12 = lm(log(salary) ~ average + runs + rbis + sos + freeagent +
                           arbitration + runsperso + hitsperso + soserrors + sbsobp + sbsruns, baseball)
  sub13 = lm(log(salary) ~ obp + runs + triples + rbis + sos + freeagent +
                           arbitration + runsperso + hitsperso + soserrors + sbsobp + sbsruns, baseball)
  sub14 = lm(log(salary) ~ average + runs + triples + rbis + sos + freeagent +
                           arbitration + runsperso + hitsperso + hrsperso + soserrors + sbsobp + sbsruns, baseball)

  extractAIC(sub12)
  extractAIC(sub13)
  extractAIC(sub14)

  ggplot(filter(allsubsinfo, p>=4), aes(p, Cp)) + geom_line(size=1, color="grey") + geom_point(size=2) +
    geom_vline(aes(xintercept=13), linetype="dashed") + geom_hline(aes(yintercept=5.196692), linetype="dashed") +
    theme_test() + theme(aspect.ratio=3/4, axis.title = element_text(size=12))
```