데이터분석방법론 기말보고서

202485-010007, 김동현

[메이저리그 야구선수 연봉 데이터]
Hitters 데이터는 263명의 야구선수에 대한 20가지 기록이다.

우선 문제를 풀기 전에 데이터를 살펴보자.
주어진 Hitters 데이터는 크게 3가지 유형으로 나눌 수 있다.
첫째는 해당 연도의 기록(AtBat, Hits, HmRun, Runs, RBI, Walks, PutOuts, Assists, Errors, Salary)이고, 둘째는 누적 기록(CAtBat, CHits, CHmRun, Runs, CRBI, CWalks)이고, 세 번째는 경력 및 소속에 대한 기록이다(Years, League, Division, NewLeague)

이 중 Salary 데이터를 종속변수로 보고, 나머지 데이터를 독립변수로 본다.
독립변수 중에서 League, Division, NewLeague 데이터는 범주형 데이터이다.
N/A는 각각 National League, American League를 의미하고, E/W는 각각 East Division, West Division을 의미한다.

문제1) PRESS, Mallow's Cp, AIC, BIC 기준 최적의 모형을 전진선택법, 후진제거법, 전체탐색법 방법을 이용하여 찾아보아라. 해당 기준들에 의하여 선택된 모형을 비교하여 보고, 본인의 최적 모형을 적절한 근거와 함께 제시하여라.

문제에서 주어진대로 4가지의 기준에 대해 3가지 방법을 사용하면 총 12가지의 모형을 도출할 수 있다.

< 기준 및 선택방법에 따른 변수 선택 결과 및 계수 >

변수	Fw_Press	Fw_Mallows_Cp	Fw_AIC	Fw_BIC	Bw_Press	Bw_Mallows_Cp	Bw_AIC	Bw_BIC	Ex_Press	Ex_Mallows_Cp	Ex_AIC	Ex_BIC
(Intercept)	91.512	117.152	162.535	91.512	162.535	117.152	162.535	117.152	117.152	130.969	118.462	-21.213
AtBat	-1.869	-2.034	-2.169	-1.869	-2.169	-2.034	-2.169	-2.034	-2.034	-2.173	-1.718	-1.782
Hits	7.604	6.855	6.912	7.604	6.918	6.855	6.918	6.855	6.855	7.358	7.613	8.491
HmRun	-	-	-	-	-	-	-	-	-	-	1.899	4.047
Runs	-	-	-	-	-	-	-	-	-	-	-	-
RBI	-	-	-	-	-	-	-	-	-	-	-	-
Walks	3.698	6.441	5.773	3.698	5.773	6.441	5.773	6.441	6.441	6.004	-	-
Years	-	-	-	-	-	-	-	-	-	-	-21.172	7.626
CAtBat	-	-	-0.130	-	-0.130	-	-0.130	-	-	-	-	-
CHits	-	-	-	-	-	-	-	-	-	-	0.121	-
CHmRun	-	-	-	-	-	-	-	-	-	1.234	-	-
CRuns	-	0.705	1.408	-	1.408	0.705	1.408	0.705	0.705	0.965	-	-
CRBI	0.643	0.527	0.774	0.643	0.774	0.527	0.774	0.527	0.527	-	0.641	-
CWalks	-	-0.807	-0.831	-	-0.831	-0.807	-0.831	-0.807	-0.807	-0.832	0.128	0.628
LeagueN	-	-	-	-	-	-	-	-	-	-	38.251	-
DivisionW	-122.952	-123.780	-112.380	-122.952	-112.380	-123.780	-112.380	-123.780	-123.780	-117.966	-	-
PutOuts	0.264	0.275	0.297	0.264	0.297	0.275	0.297	0.275	0.275	0.291	0.289	0.302
Assists	-	-	0.283	-	0.283	-	0.283	-	-	-	0.243	0.057
Errors	-	-	-	-	-	-	-	-	-	-	-4.680	-
NewLeagueN	-	-	-	-	-	-	-	-	-	-	-	-

* Fw: 전진선택법, Bw: 후진선택법, Ex: 전체탐색법

결과를 살펴보면 절반 이상의 모델에서 AtBat, Hits, Walks, CRuns, CRBI, CWalks, Division, PutOuts 변수가 선택되었다. 이 변수들을 선택한 모델은 총 4가지인데, Mallow's Cp 기준 전진선택법과 후진제거법, BIC 기준 후진제거법, PRESS 기준 전체탐색법이다.

총 12개의 모델 중에서 4개의 모델이 선택하였으므로, 대표 모델로 선택할 수 있다고 보았다.

적합된 모델은 다음과 같다.

Salary = 117.152 - 2.034 * AtBat + 6.855 * Hits + 6.441 * Walks + 0.705 * CRuns
+ 0.527 * CRBI - 0.807 * CWalks - 123.780 * DivisionW + 0.275 * PutOuts

2. 1번 문제에서 선택한 모형이 (1)정규성을 따르는지, (2)오차항의 등분산성 가정을 만족하는지 검증해보아라. 또한 (3)이상점이나 영향점이 있는지 조사하여 보아라.

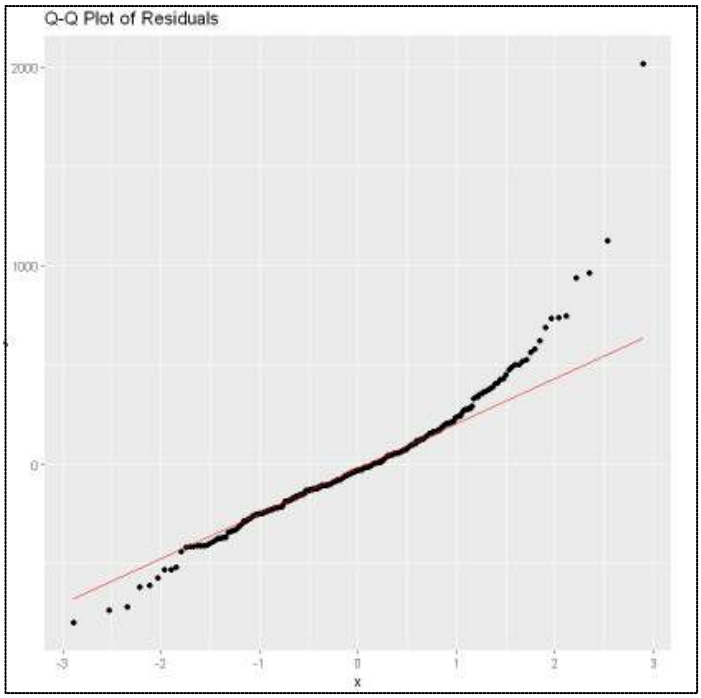
(1) 정규성을 판단하기 위해 잔차에 대해 샤피로-윌크 테스트를 실시해볼 수 있다.

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.91571, p-value = 5.075e-11
```

테스트 결과 p-value가 매우 작으므로 정규성을 충족한다고 볼 수 있다.

또한, Q-Q 플롯을 통해서도 정규성을 확인할 수 있다.

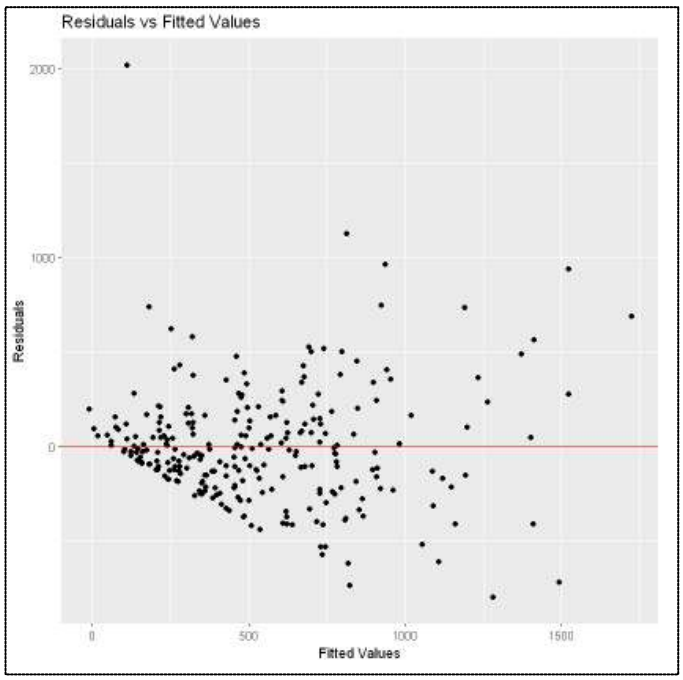


Q-Q 플롯의 모양을 보면 점들이 끝 부분에서는 다소 벗어나지만 거의 직선형태를 이루고 있어서 정규성을 충족하는 것으로 판단된다.

(2) 오차의 등분산성 가정을 만족하는지 판단하기 위해서 브루쉬-파간(Berusch-Pagan) 테스트를 실시할 수 있다.

잔차 플롯을 보면 오차가 등분산성을 가지는지 여부를 판단할 수 있다.

잔차가 일정한 패턴없이 무작위로 분포되어 있어야 하고, 잔차의 범위가 일정하게 유지되어야 하며, 특정한 형태를 보이지 않는다면 오차의 등분산성 가정이 성립한다는 의미이다.

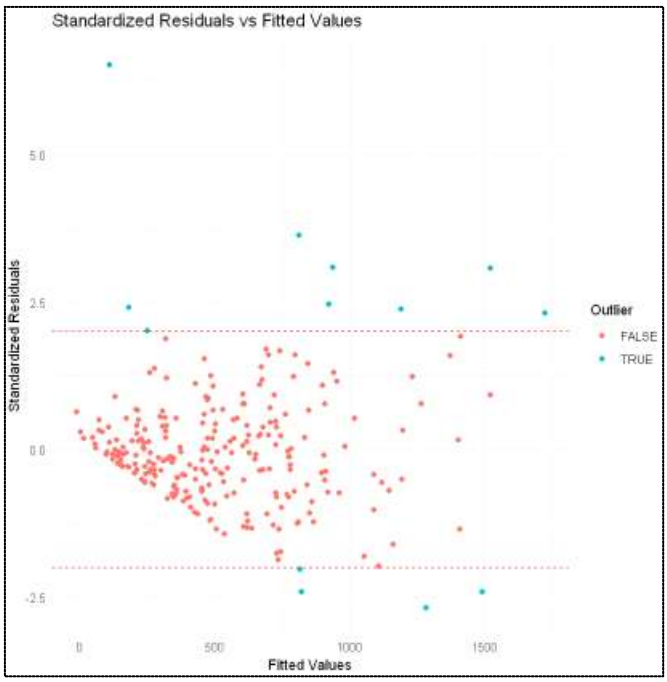


실제 그래프를 보면 y값과 잔차 사이에 일정한 경향성이 관찰되므로 오차가 등분산성을 가지지 않는 것으로 볼 수 있다.

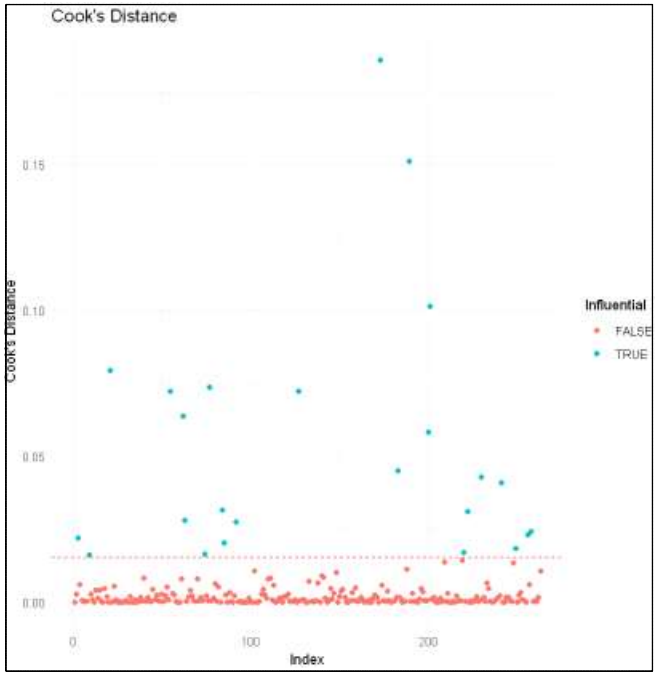
(3) 이상점이나 영향점이 있는지 조사하여 보아라.

표준화 잔차, Cook's 거리를 이용하면 이상점과 영향점을 식별할 수 있다.

표준화잔차의 절대값이 2보다 크면 이상점으로 본다고 하면, 21, 63, 77, 84, 127, 173, 183, 200, 201, 209, 222, 230, 241번 데이터가 이상점이라고 할 수 있다.



쿡의 거리가 4/n 보다 크다고 하면 영향점이라고 한다면, 3, 9, 21, 55, 62, 63, 74, 77, 84, 85, 92, 127, 173, 183, 189, 200, 201, 220, 222, 230, 241, 249, 256, 258번 데이터가 영향점이라고 할 수 있다.



3. AtBat, Hits, Walks, CAtBat, CRuns, CRBI, CWalks, League, Division, PutOuts, Assists를 설명변수로 하는 다중회귀를 적합시킨다고 할 때, 이 모형에 대한 2번 문제의 물음에 답해보아라.

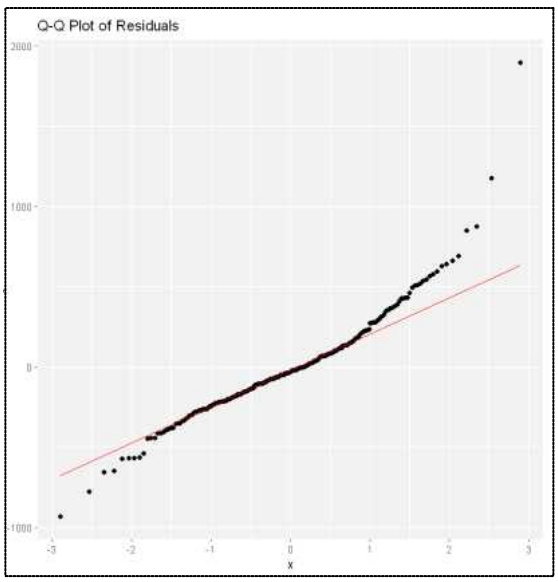
적합된 모델은 다음과 같다.

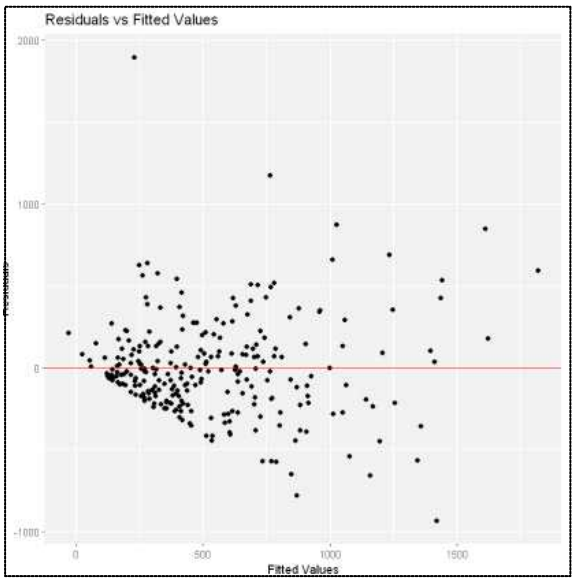
Salary = 135.751 - 2.128 * AtBat + 6.924 * Hits + 5.620 * Walks - 0.139 * CAtBat
+ 1.455 * CRuns + 0.785 * CRBI - 0.823 * CWalks + 43.112 * LeagueN
- 111.146 * DivisionW + 0.289 * PutOuts + 0.269 * Assists

샤피로-윌크 테스트와 Q-Q 플롯을 보면 새로운 모델 역시 정규성을 충족하는 것으로 보이고, 오차의 등분산성은 성립하지 않는 것으로 보인다.

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.92693, p-value = 4.354e-10
```





이상점과 영향점은 다소 달라지는 것으로 나타난다.

이상점: 3, 63, 77, 84, 127, 173, 183, 200, 201, 209, 222, 230, 241번 데이터

영향점: 3, 21, 55, 62, 63, 77, 84, 85, 92, 127, 133, 173, 183, 189, 200, 201, 209, 222, 230, 241, 249, 256, 258 번 데이터