변수 선택

회귀모델에서 독립변수의 증가는 모델의 결정계수를 증가시켜 설명력을 높이는 장점이 있지만 다중공선성 문제를 일으킬 수 있어서 추정의 신뢰도를 저하시킬 수 있고, 독립변수가 많을 경우 예측성능이 좋지 않을 가능성이 많고 독립성, 등분산성 등의 가정을 만족시키기 어렵기 때문에 독립변수를 줄일 필요가 있다.

http://contents.kocw.or.kr/KOCW/document/2015/dongguk/sonchangkyoon/7.pdf

전진 선택법(Forward Selection): 절편만 있는 모델에서 기준 통계치를 가장 많이 개선시키는 변수를 차례로 추가

후진 제거법(Backward elimination): 모든 변수가 포함된 모델에서 기준 통계치에 가장 도움이 되지 않는 변수를 하나씩 제거하는 방법

단계선택법(Stepwise selction): 모든 변수가 포함된 모델에서 출발하여 기준 통계치에 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 바져 있는 변수 중에서 기준 통계치를 가장 개선시키는 변수를 추가. 이렇게 변수의 추가 또는 삭제를 반복. 또는 절편만 포함된 모델에서 시작해 변수의 추가, 삭제를 반복할 수 도 있다.

실습1.

mlbench 패키지 안의 BostonHousing 데이터 이용
http://math.furman.edu/~dcs/courses/math47/R/library/mlbench/html/BostonHousing.html

종속변수는 medv(집의 중위가격)

1. 전진선택법 (Forward Selection)

실습1-1

```
============
# 전진선택법
library(mlbench)
data("BostonHousing")

# 회귀
ss <- lm(medv ~ .,data=BostonHousing)
# 전진선택
ss1 <- step(ss, direction = "forward")

formula(ss1)
==================
```

```
> ss1 <- step(ss, direction = "forward")
Start:  AIC=1589.64
medv ~ crim + zn + indus + chas + nox + rm + age +
 dis + rad +
    tax + ptratio + b + lstat
```

2. 후진제거법 (Backward Elimination)

다중회귀모형에서 적절한 변수 선택을 위하여 후진제거방법

실습1-2

```
# 후진제거법
==============
library(mlbench)
data("BostonHousing")

# 회귀
ss <- lm(medv ~ .,data=BostonHousing)
# 후진제거
ss2 <- step(ss, direction = "backward")

formula(ss2)
================
```

```
> # 후진제거
> ss2 <- step(ss, direction = "backward")
Start:  AIC=1589.64
medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + b + lstat

          Df Sum of Sq   RSS    AIC
- age      1      0.06 11079 1587.7
- indus    1      2.52 11081 1587.8
<none>                 11079 1589.6
- chas     1    218.97 11298 1597.5
- tax      1    242.26 11321 1598.6
- crim     1    243.22 11322 1598.6
- zn       1    257.49 11336 1599.3
- b        1    270.63 11349 1599.8
- rad      1    479.15 11558 1609.1
- nox      1    487.16 11566 1609.4
- ptratio  1   1194.23 12273 1639.4
- dis      1   1232.41 12311 1641.0
- rm       1   1871.32 12950 1666.6
- lstat    1   2410.84 13490 1687.3

Step:  AIC=1587.65
medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +
    ptratio + b + lstat

          Df Sum of Sq   RSS    AIC
- indus    1      2.52 11081 1585.8
<none>                 11079 1587.7
- chas     1    219.91 11299 1595.6
- tax      1    242.24 11321 1596.6
- crim     1    243.20 11322 1596.6
- zn       1    260.32 11339 1597.4
- b        1    272.26 11351 1597.9
- rad      1    481.09 11560 1607.2
- nox      1    520.87 11600 1608.9
- ptratio  1   1200.23 12279 1637.7
- dis      1   1352.26 12431 1643.9
- rm       1   1959.55 13038 1668.0
- lstat    1   2718.88 13798 1696.7

Step:  AIC=1585.76
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    b + lstat
```

```
Step:  AIC=1585.76
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    b + lstat

          Df Sum of Sq   RSS    AIC
<none>                  11081 1585.8
- chas     1    227.21 11309 1594.0
- crim     1    245.37 11327 1594.8
- zn       1    257.82 11339 1595.4
- b        1    270.82 11352 1596.0
- tax      1    273.62 11355 1596.1
- rad      1    500.92 11582 1606.1
- nox      1    541.91 11623 1607.9
- ptratio  1   1206.45 12288 1636.0
- dis      1   1448.94 12530 1645.9
- rm       1   1963.66 13045 1666.3
- lstat    1   2723.48 13805 1695.0
> |
```

```
> formula(ss2)
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    b + lstat
> |
```

실습2

attitude 데이터 이용
rating(등급)에 영향을 미치는 요인을 회귀를 이용해 식별
종속변수 rating에 영향을 미치는 독립변수: complaints, privileges, learning, raises, critical, advance

```
================
data(attitude)
head(attitude)

# 회귀분석
model <- lm(rating~. , data=attitude)

# 수행결과
summary(model)
=================
```

```
> data(attitude)
> head(attitude)
  rating complaints privileges learning raises critical advance
1     43         51         30       39     61       92      45
2     63         64         51       54     63       73      47
3     71         70         68       69     76       86      48
4     61         63         45       47     54       84      35
5     81         78         56       66     71       83      47
6     43         55         49       44     54       49      34
> model <- lm(rating~. , data=attitude)
>
> # 수행결과
> summary(model)

Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
complaints   0.61319    0.16098   3.809 0.000903 ***
privileges  -0.07305    0.13572  -0.538 0.595594
learning     0.32033    0.16852   1.901 0.069925 .
raises       0.08173    0.22148   0.369 0.715480
critical     0.03838    0.14700   0.261 0.796334
advance     -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05

> |
```

p-value가 <0.05 이므로 통계적으로 의미가 있음.

수정결정계수는 0.6628.

변수중 통계적으로 유의한 것은 complaints, learnings

coefficients 평가 의미

*** : 0 ~ 0.001

** : 0.001 ~ 0.01

* : 0.01 ~ 0.05

. : 0.05 ~ 0.1

```
> data(attitude)
> head(attitude)
  rating complaints privileges learning raises critical advance
1     43         51         30       39     61       92      45
2     63         64         51       54     63       73      47
3     71         70         68       69     76       86      48
4     61         63         45       47     54       84      35
5     81         78         56       66     71       83      47
6     43         55         49       44     54       49      34
> model <- lm(rating~. , data=attitude)
>
> # 수행결과
> summary(model)

Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
complaints   0.61319    0.16098   3.809 0.000903 ***
privileges  -0.07305    0.13572  -0.538 0.595594
learning     0.32033    0.16852   1.901 0.069925 .
raises       0.08173    0.22148   0.369 0.715480
critical     0.03838    0.14700   0.261 0.796334
advance     -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05

> |
```

일반적으로 사용하는 후진제거법 사용

```
====================
#독립변수 제거
reduced <- step(model, direction="backward")

summary(reduced)
==================
```

```
> reduced <- step(model, direction="backward")
Start:  AIC=123.36
rating ~ complaints + privileges + learning + raises + critical +
    advance

             Df Sum of Sq    RSS    AIC
- critical    1      3.41 1152.4 121.45
- raises      1      6.80 1155.8 121.54
- privileges  1     14.47 1163.5 121.74
- advance     1     74.11 1223.1 123.24
<none>                    1149.0 123.36
- learning    1    180.50 1329.5 125.74
- complaints  1    724.80 1873.8 136.04

Step:  AIC=121.45
rating ~ complaints + privileges + learning + raises + advance

             Df Sum of Sq    RSS    AIC
- raises      1     10.61 1163.0 119.73
- privileges  1     14.16 1166.6 119.82
- advance     1     71.27 1223.7 121.25
<none>                    1152.4 121.45
- learning    1    177.74 1330.1 123.75
- complaints  1    724.70 1877.1 134.09

Step:  AIC=119.73
rating ~ complaints + privileges + learning + advance

             Df Sum of Sq    RSS    AIC
- privileges  1     16.10 1179.1 118.14
- advance     1     61.60 1224.6 119.28
<none>                    1163.0 119.73
- learning    1    197.03 1360.0 122.42
- complaints  1   1165.94 2328.9 138.56

Step:  AIC=118.14
rating ~ complaints + learning + advance

             Df Sum of Sq    RSS    AIC
- advance     1     75.54 1254.7 118.00
<none>                    1179.1 118.14
- learning    1    186.12 1365.2 120.54
- complaints  1   1259.91 2439.0 137.94

Step:  AIC=118
rating ~ complaints + learning

             Df Sum of Sq    RSS    AIC
<none>                    1254.7 118.00
- learning    1    114.73 1369.4 118.63
- complaints  1   1370.91 2625.6 138.16
```

step에서 critical 제거 --> raise 제거 --> privileges 제거 --> advance 제거

```
> summary(reduced)

Call:
lm(formula = rating ~ complaints + learning, data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5568  -5.7331   0.6701   6.5341  10.3610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8709     7.0612   1.398    0.174
complaints    0.6435     0.1185   5.432 9.57e-06 ***
learning      0.2112     0.1344   1.571    0.128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.817 on 27 degrees of freedom
Multiple R-squared:  0.708,     Adjusted R-squared:  0.6864
F-statistic: 32.74 on 2 and 27 DF,  p-value: 6.058e-08

>
```

p-value가 < 0.05 이므로 통계적으로 의미가 있음.

수정결정계수: 0.6864

3. 단계선택법(Stepwise Selection)

실습1-3.

```
=================
# 단계적 선택방법
library(mlbench)
data("BostonHousing")

# 회귀
ss <- lm(medv ~ .,data=BostonHousing)

# 단계적선택
ss3 <- step(ss, direction = "both")
formula(ss3)

====================
```

```
> # 단계적선택
> ss3 <- step(ss, direction = "both")
Start:  AIC=1589.64
medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + b + lstat

          Df Sum of Sq   RSS    AIC
- age      1      0.06 11079 1587.7
- indus    1      2.52 11081 1587.8
<none>                 11079 1589.6
- chas     1    218.97 11298 1597.5
- tax      1    242.26 11321 1598.6
- crim     1    243.22 11322 1598.6
- zn       1    257.49 11336 1599.3
- b        1    270.63 11349 1599.8
- rad      1    479.15 11558 1609.1
- nox      1    487.16 11566 1609.4
- ptratio  1   1194.23 12273 1639.4
- dis      1   1232.41 12311 1641.0
- rm       1   1871.32 12950 1666.6
- lstat    1   2410.84 13490 1687.3

Step:  AIC=1587.65
medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +
    ptratio + b + lstat

          Df Sum of Sq   RSS    AIC
- indus    1      2.52 11081 1585.8
<none>                 11079 1587.7
+ age      1      0.06 11079 1589.6
- chas     1    219.91 11299 1595.6
- tax      1    242.24 11321 1596.6
- crim     1    243.20 11322 1596.6
- zn       1    260.32 11339 1597.4
- b        1    272.26 11351 1597.9
- rad      1    481.09 11560 1607.2
- nox      1    520.87 11600 1608.9
- ptratio  1   1200.23 12279 1637.7
- dis      1   1352.26 12431 1643.9
- rm       1   1959.55 13038 1668.0
- lstat    1   2718.88 13798 1696.7

Step:  AIC=1585.76
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    b + lstat
```

```
Step:  AIC=1585.76
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    b + lstat

          Df Sum of Sq   RSS    AIC
<none>                  11081 1585.8
+ indus    1      2.52 11079 1587.7
+ age      1      0.06 11081 1587.8
- chas     1    227.21 11309 1594.0
- crim     1    245.37 11327 1594.8
- zn       1    257.82 11339 1595.4
- b        1    270.82 11352 1596.0
- tax      1    273.62 11355 1596.1
- rad      1    500.92 11582 1606.1
- nox      1    541.91 11623 1607.9
- ptratio  1   1206.45 12288 1636.0
- dis      1   1448.94 12530 1645.9
- rm       1   1963.66 13045 1666.3
- lstat    1   2723.48 13805 1695.0
> formula(ss3)
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
    b + lstat
>
```

13개의 독립변수로 시작 (AIC값: 1,589.64) --> age변수 제거(AIC값: 1587.65) --> indus변수 제거(AIC값: 1,586.75) --> 최종 회귀식

AIC 공식

https://chukycheese.github.io/statistics/aic/