R에 의한 통계자료분석

회귀분석, 로지스틱 회귀분석

회귀분석

1. 단순선형회귀모형 적합

- 반응변수 Y와 설명변수 X 사이의 선형관계 가정

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \dots, n$$

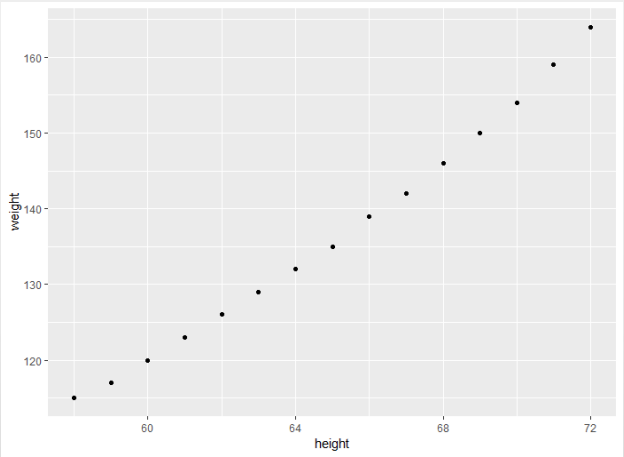
  - 일차적인 관심: 회귀계수 β_0와 β_1의 추정
  - 오차항 ε_i에 대한 가정: 서로 독립, 동일 분포 $N(0, \sigma^2)$

- 예제: 데이터 프레임 women

  - 변수 height와 weight의 관계 탐색
  - 첫 번째 작업: 산점도 작성

```
> library(ggplot2)

>  ggplot(women, aes(x=height, y=weight)) +
   geom_point()
```



선형관계가 있는 것으로 보임

- 선형회귀모형 적합: 함수 lm()

```
> fit <- lm(weight ~ height, women)

> fit

Call:
lm(formula = weight ~ height, data = women)

Coefficients:
(Intercept)       height
     -87.52         3.45
```

```
> names(fit)
 [1] "coefficients"  "residuals"   "effects"      "rank"
 [5] "fitted.values" "assign"      "qr"           "df.residual"
 [9] "xlevels"       "call"        "terms"        "model"
```

- 사용자마다 필요한 정보가 서로 다를 수 있음
- 필요한 정보를 각자 선택해서 추출
- 모든 결과를 한번에 출력하는 SAS, SPSS와는 다른 접근 방식

2. 다중선형회귀모형 적합

- 반응변수 Y와 설명변수 $X_1, X_2, \dots, X_k$ 사이에 선형 관계 가정

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \qquad i = 1, \dots, n$$

  - 오차항 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 가정: $N(0, \sigma^2)$ 서로 독립

- 선형 및 오차항 가정
  - 회귀모형의 추정 및 추론의 정당성 보장
  - 가정 위반 시 추론 결과에 대한 신뢰성 저하

- 함수 lm()의 기본적인 사용법

  lm(formula, data, subset, …)

  - formula: 회귀모형 설정을 위한 R 공식
  - data: 데이터 프레임
  - subset: 데이터의 일부분만을 이용하는 경우
    처음 100개 자료만 이용: lm(y ~ x, subset=1:100)
    변수 z가 0 이상인 자료만 이용: lm(y ~ x, subet= z>=0)

- R 공식에 사용되는 기호

1) 물결표(~): 반응변수 ~ 설명변수

2) 플러스(+): 설명변수 구분. `y ~ x1 + x2 + x3`

3) 콜론(`:`): 설명변수 사이의 상호작용. `y ~ x1 + x2 + x1:x2`

4) 점(`.`): 반응변수를 제외한 데이터 프레임에 있는 모든 변수. 데이터 프레임에 y, x1, x2, x3가 있다면 `y ~ .` → `y ~ x1 + x2 + x3`

5) 마이너스(-): 모형에서 제외되는 변수

6) `- 1` 또는 `+ 0`: 절편 제거

7) `I()`: 괄호 안의 연산자를 수학 연산자로 인식. y~I(x1+x2) → $Y = \beta_0 + \beta_1 X_1 + X_2$)

- 예제 1: 행렬 state.x77
  - 미국 50개 주와 관련된 8개 변수로 구성된 행렬
  - 반응변수: Murder

  - 행렬을 데이터 프레임으로 전환

```
> states <- as.data.frame(state.x77)
> names(states)
[1] "Population" "Income"     "Illiteracy" "Life Exp"
[5] "Murder"     "HS Grad"    "Frost"      "Area"

> states <- rename(states,
                   Life_Exp='Life Exp', Hs_Grad='HS Grad')
```

- 모형에 포함될 변수들의 관계 탐색
  - 상관계수
  - 산점도 행렬

- 상관계수 계산: 함수 cor()

```
cor(x, y=NULL, use="everything")
```

- x, y: 벡터, 행렬, 데이터 프레임
  x만 있는 경우: x에 있는 모든 변수들 사이의 상관계수 계산
  x와 y가 있는 경우: x에 있는 변수와 y에 있는 변수를 하나씩 짝을 지어 상관계수 계산
- use: 결측값 처리 방식.
  "everything": 결측값이 있으면 NA
  "pairwise": 상관계수가 계산되는 변수만을 대상으로 NA가 있는 케이스 제거

- 데이터 프레임 states에 있는 변수들의 상관계수

```
> cor(states)
            Population     Income  Illiteracy    Life_Exp
Population  1.00000000  0.2082276  0.10762237 -0.06805195
Income      0.20822756  1.0000000 -0.43707519  0.34025534
Illiteracy  0.10762237 -0.4370752  1.00000000 -0.58847793
Life Exp   -0.06805195  0.3402553 -0.58847793  1.00000000
Murder      0.34364275 -0.2300776  0.70297520 -0.78084575
HS Grad    -0.09848975  0.6199323 -0.65718861  0.58221620
Frost      -0.33215245  0.2262822 -0.67194697  0.26206801
Area        0.02254384  0.3633154  0.07726113 -0.10733194
               Murder     HS_Grad      Frost         Area
Population  0.3436428 -0.09848975 -0.3321525  0.02254384
Income     -0.2300776  0.61993232  0.2262822  0.36331544
Illiteracy  0.7029752 -0.65718861 -0.6719470  0.07726113
Life Exp   -0.7808458  0.58221620  0.2620680 -0.10733194
Murder      1.0000000 -0.48797102 -0.5388834  0.22839021
HS Grad    -0.4879710  1.00000000  0.3667797  0.33354187
Frost      -0.5388834  0.36677970  1.0000000  0.05922910
Area        0.2283902  0.33354187  0.0592291  1.00000000
```

- 상관계수 행렬: 변수의 개수가 많아지면 변수 사이 관계 파악이 어려움
- 상관계수 행렬을 그래프로 표현: 패키지 GGally의 함수 ggcorr()

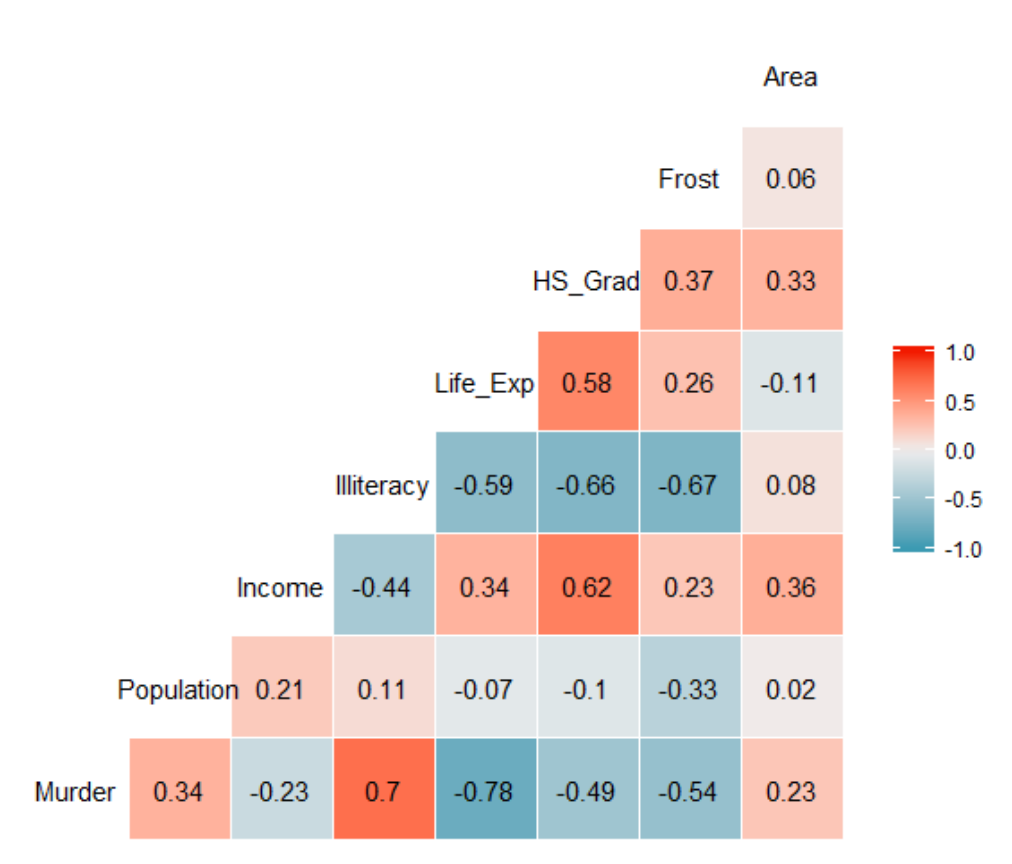
- 패키지 GGally의 함수 ggcorr()

```
ggcorr(data, method=c("pairwise", pearson"), label=FALSE,
               label_round=1, …)
```

  - label: 그래프에 상관계수 표시 여부
  - label_round: 상관계수 반올림 자릿수

- states 변수들의 상관계수 그래프

```
> library(GGally)
> states <- select(states, Murder, everything())
> ggcorr(states, label=TRUE, label_round=2)
```



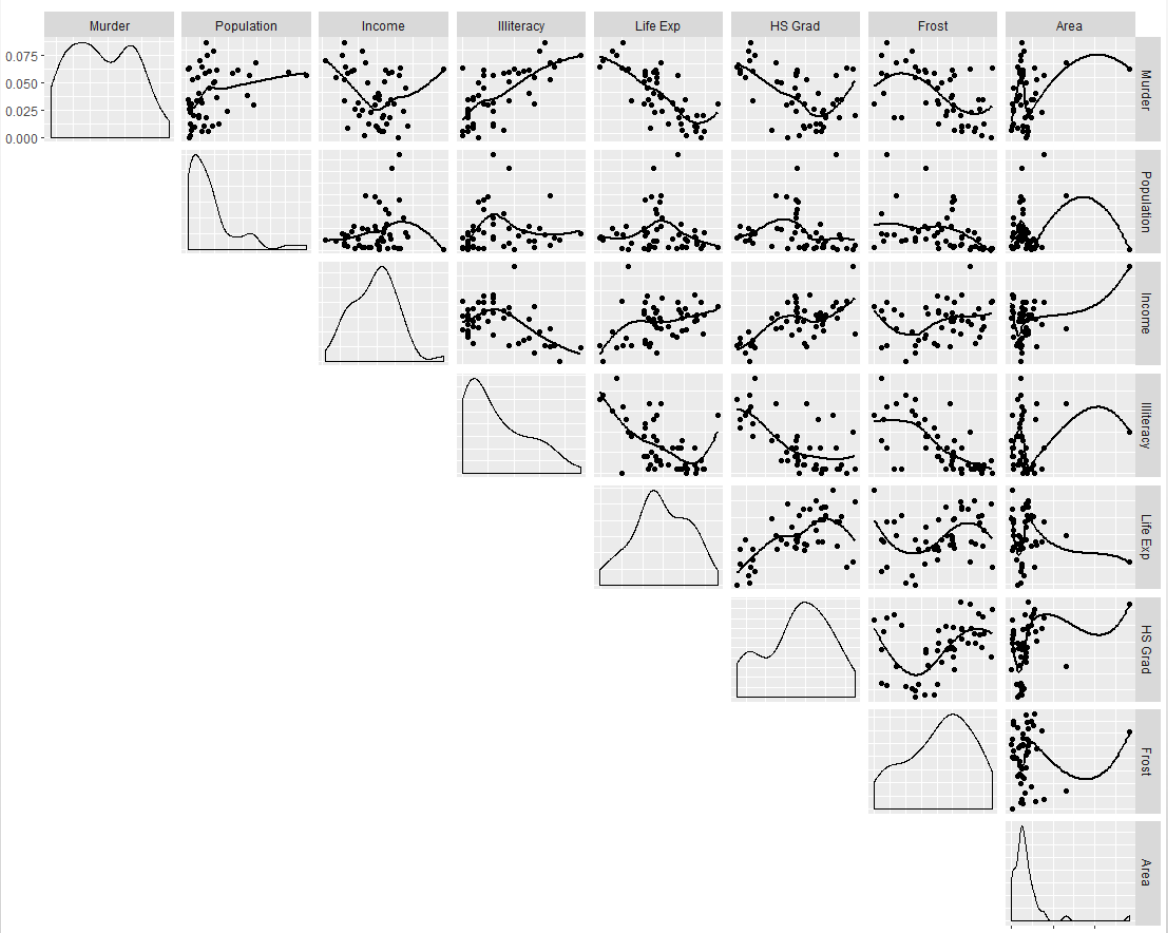
- 산점도 행렬

  - 여러 변수로 이루어진 자료에서 두 변수끼리 짝을 지어 작성된 산점도를 행렬 형태로 배열

  - 회귀분석에서 필수적인 그래프

- 패키지 GGally의 함수 ggpairs()

```
> library(GGally)
> ggpairs(states, lower=list(continuous="blank"),
          upper=list(continuous=wrap("smooth_loess", se=FALSE)))
```



- 예제 1 계속: states에 대한 회귀모형 적합

```
> fit <- lm(Murder ~ ., states)

> fit

Call:
lm(formula = Murder ~ ., data = states)

Coefficients:
(Intercept)   Population       Income   Illiteracy
  1.222e+02    1.880e-04   -1.592e-04    1.373e+00
   Life_Exp      Hs_Grad        Frost          Area
 -1.655e+00    3.234e-02   -1.288e-02    5.967e-06
```

- 함수 lm()으로 생성된 객체(회귀분석 결과)의 내용 확인을 위한 함수
  - anova(): 분산분석표
  - coefficients(): 추정된 회귀계수, coef()도 가능
  - confint(): 회귀계수 신뢰구간.
  - fitted(): 반응변수 적합값
  - residuals(): 잔차. resid()도 가능
  - summary(): 중요한 적합 결과 요약

- 예제 2: women

  - 데이터 프레임 women의 변수 weight와 height의 관계
  - 선형보다는 2차가 더 적합한 것으로 보임

  - 다항회귀모형

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2{}_i + \cdots + \beta_p X^p{}_i + \varepsilon_i$$

    - 차수 p를 너무 높이면 다중공선성의 문제가 발생할 수 있음

    - 3차를 넘지 않는 것이 일반적

- 반응변수 weight에 대한 height의 2차 다항회귀모형 적합

```
> fit_w <- lm(weight ~ height + I(height^2), women)

> fit_w

Call:
lm(formula = weight ~ height + I(height^2), data = women)

Coefficients:
(Intercept)       height  I(height^2)
  261.87818     -7.34832      0.08306
```

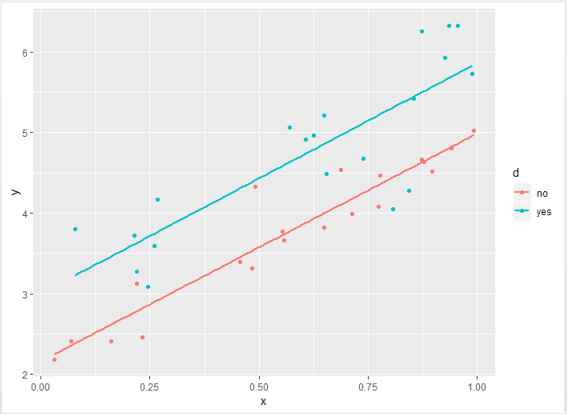
모형식: $\hat{Y}_i = 261.87 - 7.34X_i + 0.083X_i^2$

- 예제 3: 질적 변수를 설명변수로 사용

  - 회귀모형에서 사용되는 변수 형태
    반응변수: 연속형(정규분포 가정 필요)
    설명변수: 연속형(정규분포 가정은 필요 없으나, 가능한 좌우대칭)
    범주형(가변수 필요)

  - 가변수 회귀모형: 2개 범주(yes, no) → 1개 가변수

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon, \quad D = \begin{cases} 0 & no \\ 1 & yes \end{cases}$$



  - D=0인 범주: 기준 범주
  - 회귀계수 β_2: yes 범주와 기준 범주의 차이

  - 일반적으로 가변수 개수=범주 개수-1

  - 만일 가변수 개수=범주 개수이면 회귀계수 추정이 불가능
    → 절편 제거하면 추정 가능
    → 회귀계수의 해석이 달라짐(해당 범주의 효과)
    → 두 개 이상의 범주형 변수가 포함되는 경우에는 적용이 어려움

- 패키지 carData의 데이터 프레임 Leinhardt
  - 1970년대 105개 나라의 신생아 사망률, 소득, 지역 및 원유 수출 여부
  - 반응변수: 신생아 사망률(infant)
  - 설명변수: 소득(income), 지역(region, 4개 수준: Africa, Americas, Asia, Europe), 원유 수출(oil, 2개 수준: no, yes)

- 함수 lm()에 요인 입력: 자동으로 필요한 개수의 가변수 포함

```
>lm(infant ~ income + region, data=Leinhardt)

Call:
lm(formula = infant ~ income + region, data = Leinhardt)

Coefficients:
   (Intercept)          income  regionAmericas       regionAsia
     1.432e+02      -3.458e-03      -8.473e+01      -4.480e+01
  regionEurope
    -1.135e+02
```

  - 기준 범주: 알파벳 첫 번째 범주인 Africa
  - 회귀계수 regionAmericas는 범주 Americas와 기준 범주 Africa의 차이

- 절편 제거 모형: + 0 또는 -1 포함

```
> lm(infant ~ income + region + 0, data=Leinhardt)

Call:
lm(formula = infant ~ income + region + 0, data = Leinhardt)

Coefficients:
        income    regionAfrica  regionAmericas       regionAsia
     -0.003458       143.235952        58.504549        98.440309
  regionEurope
     29.767837
```

- 두 범주형 변수(region, oil) 포함

```
> lm(infant ~ income + region + oil, data=Leinhardt)

Call:
lm(formula = infant ~ income + region + oil, data = Leinhardt)

Coefficients:
   (Intercept)          income  regionAmericas       regionAsia
     136.82468        -0.00529       -83.64943        -45.88540
  regionEurope          oilyes
    -101.48624         78.33508
```

3. 회귀모형의 추론

- 회귀모형: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$

- 회귀계수에 대한 가설
  1) $H_0: \beta_1 = \cdots = \beta_k = 0$
  2) $H_0: \beta_q = \beta_{q+1} = \cdots = \beta_r = 0, \quad q < r \leq k$
  3) $H_0: \beta_i = 0, H_1: \beta_i \neq 0$

- 회귀계수의 신뢰구간

- 회귀모형 적합 정도에 대한 통계량
  - 결정계수, 수정된 결정계수
  - AIC, BIC

- 적합된 회귀모형 추론을 위한 함수

  - 함수 summary()

```
> fit <- lm(Murder ~ ., states)
> summary(fit)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4452 -1.1016 -0.0598  1.1758  3.2355

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.222e+02  1.789e+01   6.831 2.54e-08 ***
Population   1.880e-04  6.474e-05   2.905  0.00584 **
Income      -1.592e-04  5.725e-04  -0.278  0.78232
Illiteracy   1.373e+00  8.322e-01   1.650  0.10641
Life_Exp    -1.655e+00  2.562e-01  -6.459 8.68e-08 ***
Hs_Grad      3.234e-02  5.725e-02   0.565  0.57519
Frost       -1.288e-02  7.392e-03  -1.743  0.08867 .
Area         5.967e-06  3.801e-06   1.570  0.12391
---

Residual standard error: 1.746 on 42 degrees of freedom
Multiple R-squared:  0.8083,      Adjusted R-squared:
0.7763
F-statistic: 25.29 on 7 and 42 DF,  p-value: 3.872e-13
```

- 개별 회귀계수 추정 및 검정
- $\sqrt{MSE}$
- 결정계수 및 수정된 결정계수
- 모든 회귀계수의 유의성 검정

- 두 회귀모형의 비교

1) 확장모형(Ω): $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$

2) 축소모형(ω): 다음의 귀무가설이 사실인 모형

$$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_r = 0, \qquad q < r \leq k$$

RSS_Ω: 확장모형의 잔차제곱합
RSS_ω: 축소모형의 잔차제곱합

- 만일 $RSS_\omega - RSS_\Omega$가 적다면, 축소모형이 확장모형만큼 좋다는 의미
- 모수절약의 원칙에 따라 축소모형 선택 가능

- 검정통계량

$$F = \frac{(RSS_\omega - RSS_\Omega) \text{두 모형의 모수 차이}}{RSS_\Omega \, n - k - 1}$$

- 함수 anova()

  - 두 회귀모형의 비교

```
> fit <- lm(Murder ~ ., states)

> fit1 <- lm(Murder ~ Population + Illiteracy + Life_Exp, states)

> anova(fit1, fit)
Analysis of Variance Table

Model 1: Murder ~ Population + Illiteracy + Life_Exp
Model 2: Murder ~ Population + Income + Illiteracy + Life_Exp + Hs_Grad +
    Frost + Area
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 152.49
2     42 128.03  4    24.461 2.0061 0.1111
```

anova(축소모형, 확장모형)

귀무가설의 기각이 어려움

- 회귀모형 적합 정도에 대한 통계량

  - 결정계수(R^2): 반응변수의 변량 중 회귀모형으로 설명되는 변량의 비율
    - 모형에 포함된 설명변수의 개수가 증가하면 증가하는 특성이 있음
    - 설명변수의 개수가 같은 모형 비교에는 의미가 있는 통계량

  - 수정 결정계수(adj. R^2): 추가된 설명변수가 모형 적합도에 도움이 되는 경우에만 증가

  - AIC & BIC: 설명변수의 개수가 p인 모형
    - $AIC = n\log\left(\frac{SSE}{n}\right) + 2p$
    - $BIC = n\log\left(\frac{SSE}{n}\right) + p\log(n)$

    - AIC, BIC가 작은 모형이 더 적합도가 높은 모형

- R에서 모형의 적합도 계산

```
> fit <- lm(Murder ~ ., states)

>summary(fit)$r.squared
[1] 0.8082607

> summary(fit)$adj.r.squared
[1] 0.7763042

>AIC(fit)  [
1] 206.9071

> BIC(fit)
[1] 224.1153
```

4. 변수 선택

- 반응변수의 변동을 설명할 수 있는 많은 설명변수 중 '최적'의 변수를 선택하여 모형에 포함시키는 절차

- 검정에 의한 방법
  - 변수의 유의성 검정을 이용하여 단계적으로 모형 선택
  - 후진소거법, 전진선택법, 단계별 선택법

- 모형선택 기준에 의한 방법
  - 모형의 적합도 등을 측정하는 통계량을 기반으로 모형 선택
  - 결정계수, 수정결정계수, 잔차제곱합, C_p 통계량, AIC, BIC 등등

- 어떤 모형이 '최적' 모형인가?

모형선택 기준에 의한 방법

- 모형 수립 목적을 고려한 변수 선택 방법
- 모형의 적합도 등을 나타내는 통계량을 선택 기준으로 사용
- 사용되는 통계량
  - 수정 결정계수(adj. R^2)
  - AIC, BIC
- 선택 방법
  - 모든 가능한 회귀(All possible regression)
  - 단계별 선택법

- 모든 가능한 회귀
  - 설명변수의 모든 가능한 조합에 대하여 선택 기준으로 사용되는 통계량 계산
  - 특정 통계량을 기준으로 가장 최적인 모형을 보여주는 방식
  - leaps::regsubsets()로 실시

- 함수 regsubsets( )으로 적합

```
> library(leaps)
> fits <- regsubsets(Murder ~ ., states)
```

- 설명변수가 k인 모형 중 결정계수가 가장 높은 모형

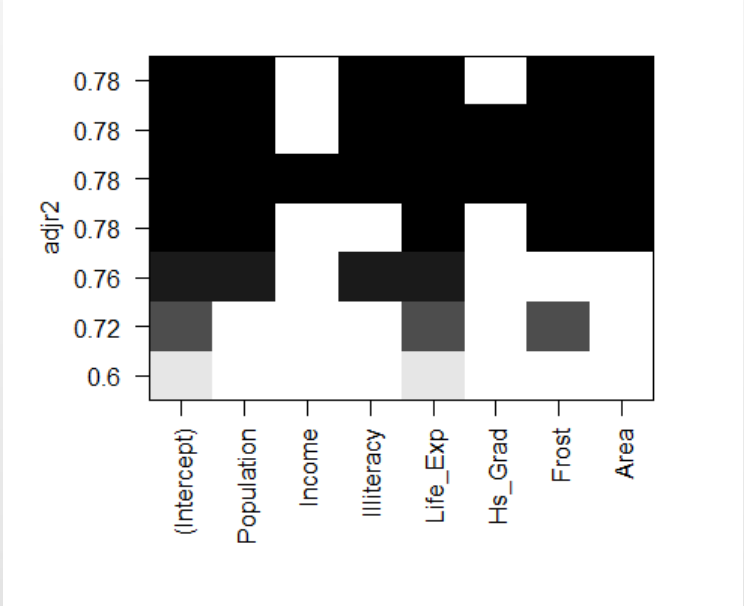
```
> summary(fits)
         Population Income Illiteracy Life_Exp Hs_Grad Frost Area
1  ( 1 ) " "        " "    " "        "*"      " "     " "   " "
2  ( 1 ) " "        " "    " "        "*"      " "     "*"   " "
3  ( 1 ) "*"        " "    "*"        "*"      " "     " "   " "
4  ( 1 ) "*"        " "    " "        "*"      " "     "*"   "*"
5  ( 1 ) "*"        " "    "*"        "*"      " "     "*"   "*"
6  ( 1 ) "*"        " "    "*"        "*"      "*"     "*"   "*"
7  ( 1 ) "*"        "*"    "*"        "*"      "*"     "*"   "*"
```

- 모든 가능한 회귀의 적합 결과 확인

1) 함수 plot( )에 의한 확인

```
> plot(fits, scale="adjr2")
```

scale: 디폴트 "bic"



- 각 행: 하나의 모형을 의미
- 색이 채워진 직사각형: 모형에 포함된 변수
- Y축: 각 모형의 adj. R^2 값
- 위에서 첫 번째 모형: adj. R^2 의 값이 가장 큰 모형
- Population, Illteracy, Life_Exp, Frost, Area 포함 모형 선택

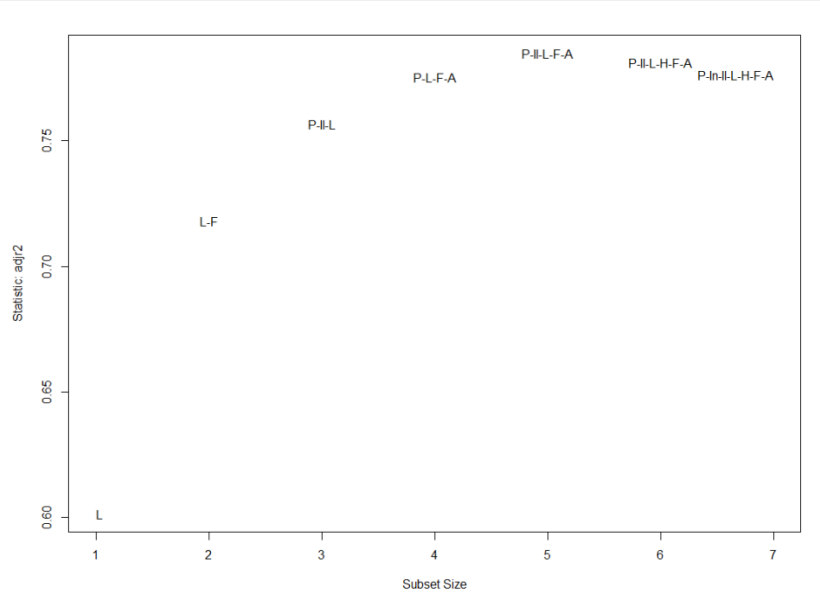
2) 함수 car::subsets( )에 의한 확인

```
subsets(lm object,
              statistics=c(“bic”, “cp”, “adjr2”, “rsq”, “rss”),
              legend=“interactive”)
```

- 옵션 statistics: 디폴트 bic

- 옵션 legend: 범례의 위치
  legend=”interactive”: 디폴트, 마우스로 위치 지정 가능
  legend=FALSE: Console 창에 범례 출력

```
>  subsets(fits, statistic="adjr2", legend=FALSE)
   Abbreviation
Population              P
Income                 In
Illiteracy             Il
Life_Exp                L
Hs_Grad                 H
Frost                   F
Area                    A
```



- P-Il-L-F-A 모형 선택

- 단계별 선택

  - AIC 혹은 BIC에 의한 단계별 선택
  - 변수의 개수가 많은 경우에 적용 가능
  - MASS::stepAIC( )로 실시

    stepAIC(object, scope, k=2)

    - object: 함수 glm( )으로 생성된 객체
    - scope: 모든 설명변수가 포함된 full 모형의 formula. 생략되면 object에 설정된 모형이 full 모형.
    - k: 탐색에 사용되는 IC. k=2는 AIC, k=log(n)은 BIC.

- 전진선택법에 의한 단계별 선택

```
> library(MASS)

> fit_full <- lm(Murder ~ ., states)

> fit <- lm(Murder ~ 1, states)

> stepAIC(fit, scope=formula(fit_full))
```

```
> formula(fit_full)
Murder ~ Population + Income + Illiteracy + Life_Exp + Hs_Grad +
    Frost + Area
```

```
Start:  AIC=131.59
Murder ~ 1

             Df Sum of Sq    RSS     AIC
+ Life_Exp    1    407.14 260.61  86.550
+ Illiteracy  1    329.98 337.76  99.516
+ Frost       1    193.91 473.84 116.442
+ Hs_Grad     1    159.00 508.75 119.996
+ Population  1     78.85 588.89 127.311
+ Income      1     35.35 632.40 130.875
+ Area        1     34.83 632.91 130.916
<none>                   667.75 131.594

Step:  AIC=86.55
Murder ~ Life_Exp

             Df Sum of Sq    RSS     AIC
+ Frost       1     80.10 180.50  70.187
+ Illiteracy  1     60.55 200.06  75.329
+ Population  1     56.62 203.99  76.303
+ Area        1     14.12 246.49  85.764
<none>                   260.61  86.550
+ Hs_Grad     1      1.12 259.48  88.334
+ Income      1      0.96 259.65  88.366
- Life_Exp    1    407.14 667.75 131.594

Step:  AIC=70.19
Murder ~ Life_Exp + Frost

             Df Sum of Sq    RSS     AIC
+ Population  1    23.710 156.79  65.146
+ Area        1    21.084 159.42  65.976
<none>                   180.50  70.187
+ Illiteracy  1     6.066 174.44  70.477
+ Income      1     5.560 174.94  70.622
+ Hs_Grad     1     2.068 178.44  71.610
- Frost       1    80.104 260.61  86.550
- Life_Exp    1   293.331 473.84 116.442
```

```
Step:  AIC=65.15
Murder ~ Life_Exp + Frost + Population

             Df Sum of Sq    RSS     AIC
+ Area        1    19.040 137.75  60.672
+ Illiteracy  1    11.826 144.97  63.225
<none>                   156.79  65.146
+ Hs_Grad     1     1.821 154.97  66.561
+ Income      1     0.739 156.06  66.909
- Population  1    23.710 180.50  70.187
- Frost       1    47.198 203.99  76.303
- Life_Exp    1   296.694 453.49 116.247

Step:  AIC=60.67
Murder ~ Life_Exp + Frost + Population + Area

             Df Sum of Sq    RSS     AIC
+ Illiteracy  1     8.723 129.03  59.402
<none>                   137.75  60.672
+ Income      1     1.241 136.51  62.220
+ Hs_Grad     1     0.771 136.98  62.392
- Area        1    19.040 156.79  65.146
- Population  1    21.666 159.42  65.976
- Frost       1    52.970 190.72  74.940
- Life_Exp    1   272.927 410.68 113.290

Step:  AIC=59.4
Murder ~ Life_Exp + Frost + Population + Area + Illiteracy

             Df Sum of Sq    RSS    AIC
<none>                   129.03 59.402
- Illiteracy  1     8.723 137.75 60.672
+ Hs_Grad     1     0.763 128.27 61.105
+ Income      1     0.026 129.01 61.392
- Frost       1    11.030 140.06 61.503
- Area        1    15.937 144.97 63.225
- Population  1    26.415 155.45 66.714
- Life_Exp    1   140.391 269.42 94.213
```

- 후진소거법에 의한 단계별 선택

```
> stepAIC(fit_full, trace=FALSE)

Call:
lm(formula = Murder ~ Population + Illiteracy + Life_Exp + Frost +
    Area, data = states)

Coefficients:
(Intercept)   Population   Illiteracy     Life_Exp        Frost
  1.202e+02    1.780e-04    1.173e+00   -1.608e+00   -1.373e-02
       Area
  6.804e-06
```

- BIC에 의한 단계별 선택

```
> stepAIC(fit_full, k=log(nrow(states)), trace=FALSE)

Call:
lm(formula = Murder ~ Population + Life_Exp + Frost + Area, data
= states)

Coefficients:  (Intercep
t)            Population     Life_Exp         Frost          Area
  1.387e+02    1.581e-04   -1.837e+00   -2.204e-02    7.387e-06
```

AIC에 의한 단계별 선택과는 다른 결과

5. 회귀진단

- 회귀진단
  1) 회귀모형에 대한 진단
  2) 관찰값에 대한 진단

- 회귀모형에 대한 진단
  - 회귀모형의 가정 사항 만족 여부 확인
  - 적합 및 추론 결과의 신빙성 확보

- 관찰값에 대한 진단
  - 개별 관찰값이 모형 추정 과정에 미치는 영향력 파악