4 장 Multiple Linear Regression

CONTENTS

4.1 서론

- 다중선형회귀는 스칼라 반응변수(반응변수가 1개)와 두 개 이상의 설명변수 간의 관계를 모 형화 하는 방법이다. 설명변수가 한 개인 경우는 단순 선형 회귀에 해당한다. 여러 개의 상관된 반응변수를 다루는 다변량 다중회귀와는 용어의 구분이 필요하다.
- 이 장에서는 다중선형회귀분석의 전반에 대한 주요 내용을 다룬다. 여기에는 다중회귀의 원리, 다중회귀의 적합 과정과 변수선택법등을 다룬다.

4.2 다중선형회귀

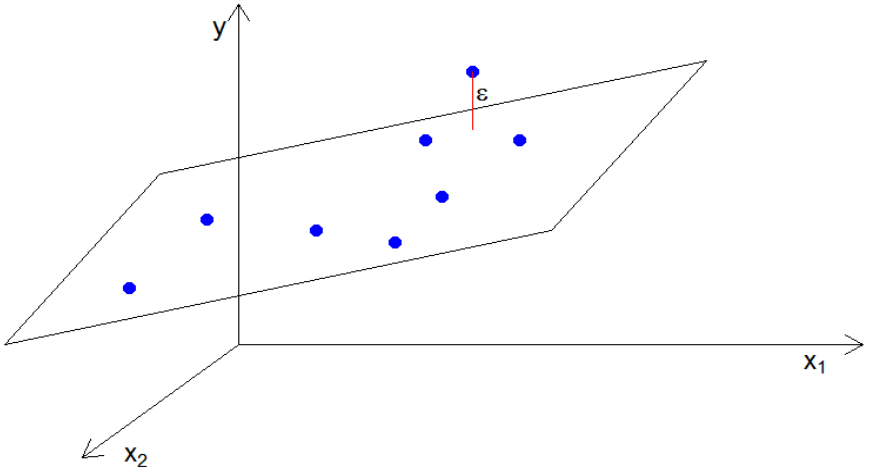
- 다중선형회귀(multiple linear regression)는 단순선형회귀의 확장으로 독립변수의 수가 여러 개인 경우에 해당하며, 모형은 다음과 같다. 오차(ϵ)에 대한 가정은 단순선형회귀의 경우와 동일하다.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon.$$

4.2 다중선형회귀

4.2.1 다중선형회귀의 원리

- 다중선형회귀는 아래의 [그림 4.1]에서와 같이 관측값 y와 미지인 초평면(hyperplane) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$까지의 ($y$축 방향으로의) 수직거리의 제곱합 즉, 오차(ϵ)들의 제곱합 $\sum_{i=1}^{n} \epsilon_i^2$을 최소로 하는 모수($b_0, b_1, \ldots, b_p$)를 찾는 것이다.



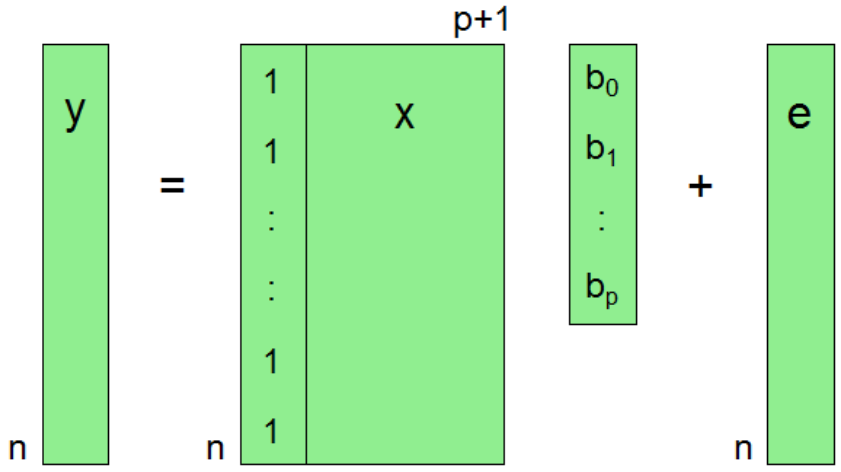
[그림 4.1] 다중회귀의 원리

4.2 다중선형회귀

- 최소제곱법 또는 OLS(Ordinary least square)에 의해 회귀계수($\beta_0, \beta_1, \cdots, \beta_p$)는 다음 식

$$b = (X\ X)^{-1} X\ y$$

으로부터 추정된다. 위 식에서 X, y에 대한 정의는 다음 [그림 4.2]와 같다.



[그림 4.2] 다중회귀의 모형식

4.2 다중선형회귀

4.2.2 다중상관계수

- 다중상관계수(multiple correlation coefficient, multiple R^2)는 다중회귀에서 반응변수에 대한 예측력을 평가하는데 사용된다. 이 값은 예측값(추정값)과 관측값 사이의 상관계수의 제곱(제곱상관계수)에 해당한다.
- 또한, 이 값은 독립변수에 의해 설명되어지는 반응변수의 분산의 비율로도 해석할 수 있다. 다중상관계수는 총제곱합(SST)에서 회귀제곱합(SSR)이 차지하는 비율로 다음과 같이 정의된다.

$$R^2 = \frac{SSR}{SST}$$

- 다중상관계수의 통계적 유의성은 F-검정을 이용한다.

4.2 다중선형회귀

유 의 다중상관계수는 그 의미상 제곱다중상관계수(squared multiple correlation coefficient)의 표현이 더 정확하나, 통상적으로 "제곱"의 표현을 생략하고 사용한다.

- 다중상관계수는 모상관계수를 과대추정(overestimate)하는 경향이 있다. 수정상관계수 (adjusted R^2)는 이를 보완한 추정량으로 다음과 같이 정의된다.

$$\text{Adjusted } R^2 = 1 - \frac{SS_{residuals} / (n-K)}{SS_{total} / (n-1)}$$

위 식에서 n 은 표본의 크기, K 는 예측변수의 수를 나타낸다.

4.2 다중선형회귀

예제 1 Prestige{car} 자료를 이용하여 상관분석과 다중회귀분석을 수행한다.

```
> library(car)
> data(Prestige)
> str(Prestige)
'data.frame':  102 obs. of  6 variables:
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377  11023 ...
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8  ...
 $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153  ...
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2  2 ...
```

4.2 다중선형회귀

```
> head(Prestige)     education  income  women  prestige  census  type

gov.administrators       13.11   12351  11.16      68.8    1113  prof
general.managers         12.26   25879   4.02      69.1    1130  prof
accountants              12.77    9271  15.70      63.4    1171  prof
purchasing.officers      11.42    8865   9.11      56.8    1175  prof
chemists                 14.62    8403  11.68      73.5    2111  prof
physicists               15.64   11030   5.13      77.6    2113  prof
-------------------------------------
```

자료 설명

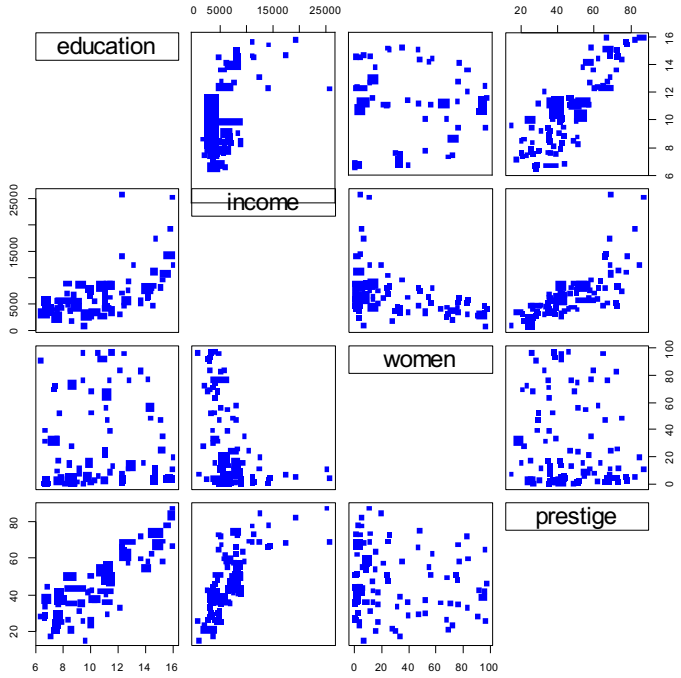
이 자료는 1971년 캐나다의 직업에 대한 자료이다. 102개의 직업군별로 6개의 변수가 조사되었다. 변수는 평균교육연수(education), 평균연봉(income), 여성비율(women), 명망점수(prestige), 인구조사직업코드(census), 직업형태(type)이다. 여기서는 income을 반응변수로 education, women, prestige을 예측변수로 하는 다중회귀모형을 적합하고자 한다.

4.2 다중선형회귀

```
> summary(Prestige)
   education          income           women          prestige    
 Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80  
 1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23  
 Median :10.540   Median : 5930   Median :13.600   Median :43.60  
 Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83  
 3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27  
 Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20  
     census       type   
 Min.   :1113   bc  :44  
 1st Qu.:3120   prof:31  
 Median :5135   wc  :23  
 Mean   :5402   NA's: 4  
 3rd Qu.:8312            
 Max.   :9517            
```

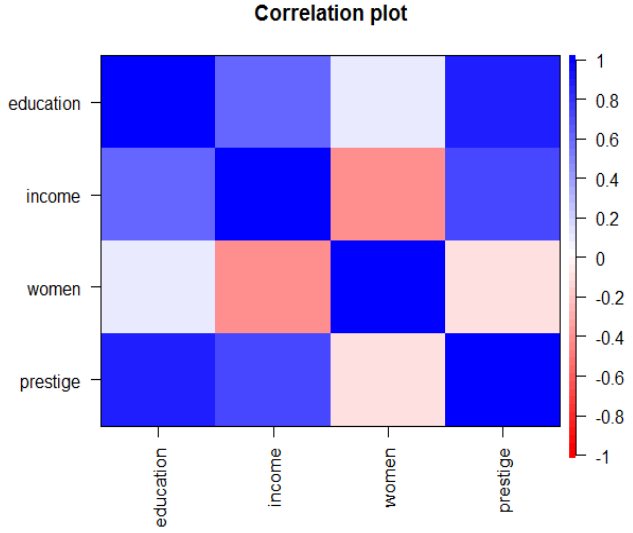
4.2 다중선형회귀

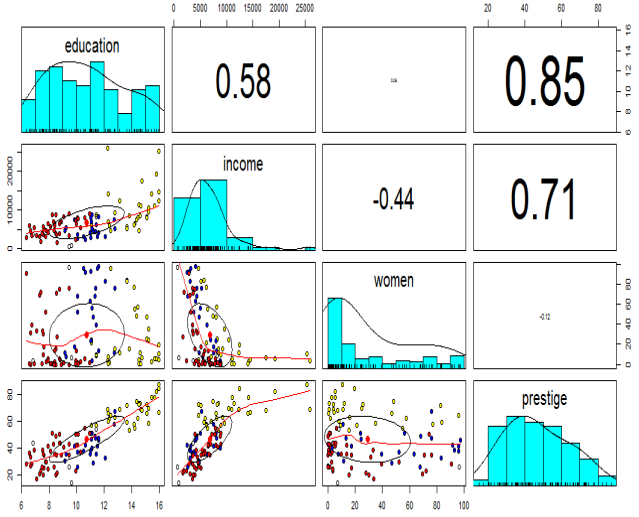
```
> # 상관분석 시각화
> plot(Prestige[,1:4], pch=15, col="blue")
```



4.2 다중선형회귀

```
> # pairs.panel{psych} 함수 이용
> library(psych)
> pairs.panels(Prestige[,1:4], scale=T)
> pairs.panels(Prestige[1:4],
      bg=c("red","yellow","blue")[Prestige$type], pch=21)
```





4.2 다중선형회귀

```
> ## 다중선형회귀 적합: lm() 함수 이용
> mod1 = lm(income ~ education + prestige + women, data=Prestige)
> summary(mod1)

Call:
lm(formula = income ~ education + prestige + women, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-7715.3  -929.7  -231.2   689.7 14391.8 
                                   (…)
```

4.2 다중선형회귀

```
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -253.850    1086.157   -0.234     0.816
education    177.199     187.632    0.944     0.347
prestige     141.435      29.910    4.729 7.58e-06 ***
women        -50.896       8.556   -5.948 4.19e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2575 on 98 degrees of freedom
Multiple R-squared:  0.6432, Adjusted R-squared:  0.6323
F-statistic: 58.89 on 3 and 98 DF,  p-value: < 2.2e-16
```

해 석 education 변수가 유의하지 않으므로, 이를 제외한 모형을 적합하기로 한다.

4.2 다중선형회귀

```
> mod2 = lm(income ~prestige + women, data=Prestige)
> summary(mod2)

Call:
lm(formula = income ~ prestige + women, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-7620.9 -1008.7  -240.4   873.1 14180.0 
                                   (…)
```

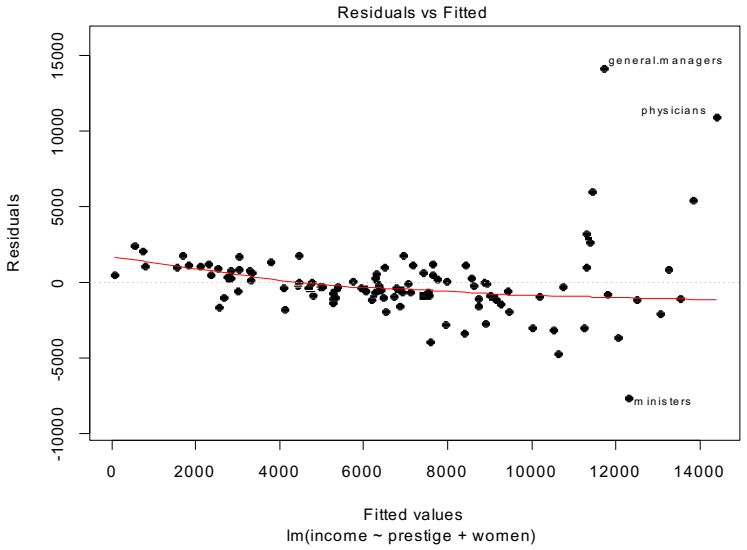
4.2 다중선형회귀

```
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   431.574    807.630   0.534     0.594
prestige      165.875     14.988  11.067   < 2e-16 ***
women         -48.385      8.128  -5.953 4.02e-08 ***
---
Signif. codes:   0 ‘***’ 0.001 ‘**’ 0.01  ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2573 on 99 degrees of freedom
Multiple R-squared:   0.64,  Adjusted R-squared:  0.6327
F-statistic: 87.98 on 2 and 99 DF,  p-value: < 2.2e-16
```

해 석 모형이 잘 적합되며, 두 변수(prestige, women)가 모두 유의하다.

4.2 다중선형회귀

```
> plot(mod2, pch=16, which=1)     # 잔차 그림
```



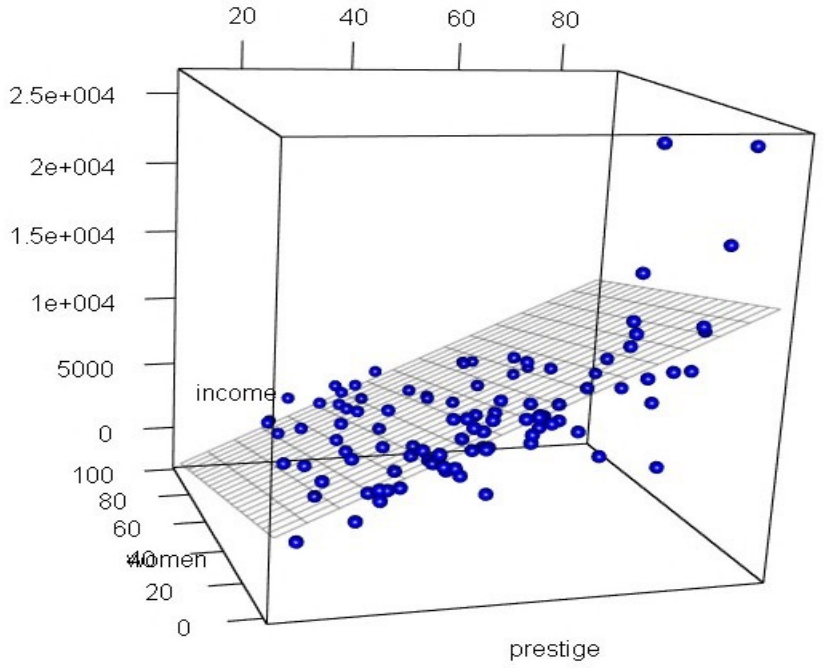
해 석

잔차분석 결과 오차에 대한 등분산성 가정이 위배되며, 비선형성을 보이고 있어 모형의 개선이 필요해 보인다. 잔차분석을 포함한 보다 자세한 회귀진단과 모형에 대한 개선은 여기서 다루지 않기로 한다.

4.2 다중선형회귀

```
> ## 다중회귀 적합결과의 시각화
> library(rgl)
> newdat <- expand.grid(prestige=seq(10,90,by=5),
      women=seq(0,100,by=5))
> newdat$pp <- predict(mod2,newdata=newdat)
> library(scatterplot3d)
> with(Prestige[, 1:4], plot3d(prestige,women,income, col="blue",
             size=1, type="s"))
> with(newdat,surface3d(unique(prestige),unique(women),pp,
                        alpha=0.3,front="line", back="line"))
```

4.2 다중선형회귀


20
40
60
80
2.5e+004
2e+004
1.5e+004
1e+004
5000
0
income
100
80
60
40
20
0
women
prestige

4.3 변수선택법

- 다중회귀에서 최종 적합모형에 포함될 예측변수의 선택은 중요하다. 변수선택의 기본 원리는 "데이터에 대한 설명력이 뛰어나며, 동시에 단순한 형태를 가지는 모형"을 찾는 것이다. 변수 선택에는 두 가지 접근법이 있다. 하나는 모든 가능한 회귀 접근법과 자동화된 변수선택법이다.

4.3 변수선택법

(a) 모든 가능한 회귀 접근법(all possible regression approach)

- 이 방법은 예측변수의 모든 가능한 부분집합에 대해 모형을 적합한 뒤, 적절한 판정기준(예를 들어, 수정결정계수, AIC, BIC)에 의해 최적의 모형을 선택하는 방법이다. 이들 기준은 각 모형에 점수를 할당하고, 가장 우수한 점수를 가지는 모형을 선택하도록 한다.

4.3 변수선택법

- R의 regsubsets{leaps} 함수는 모든 가능한 회귀를 통한 변수 선택에 유용하다. 이 함수의 적용 결과를 시각화하여 변수를 선택할 수 있다. 이 함수의 일반 형식은 다음과 같다.

```
regsubsets(x=, data=, weights=NULL, nbest=1, nvmax=8, force.in=NULL,
        force.out=NULL, intercept=TRUE, method=c("exhaustive",
        "backward", "forward", "seqrep"), really.big=FALSE,...)
```

  - x= 계획행렬 또는 모형식 지정
  - nbest= 변수의 수별로 상위 몇 개를 나타낼지를 지정
  - nvmax= 부분집합의 최댓값(변수의 수) 지정
  - method= 탐색방법을 지정

4.3 변수선택법

- 위 함수의 수행 결과 객체(regsubsets 객체)에 대해 다음의 함수가 유용하다.

```
plot(object, scale=c("adjr2", "aic", "bic", ...))
summary(object, all.best=TRUE,
matrix=TRUE,
        matrix.logical=FALSE,df=NULL,...)
```

- all.best= 모든 최적의 부분집합 도는 예측변수의 크기별로 하나의 최적모형 제공

```
coef(object,id,vcov=FALSE,...
) vcov(object,id,...)
```

4.3 변수선택법

(b) 자동화된 변수선택법

- 자동화된 변수선택법은 예측변수의 수가 크고, 따라서 모든 가능한 회귀 방법의 적용이 어려운 경우에 유용하다. 이 경우, 적절한 탐색 알고리즘(예를 들어, 전진선택법, 후진제거법, 단계별선택법)을 사용하여 최적의 모형을 찾는 것이 보다 효율적이다.
  - **전진선택법(forward selection method) :** 절편항만 포함하는 가장 작은 모형에서 반응변수에 가장 큰 영향을 주는 설명변수를 차례로 모형에 포함시켜나가되 더 이상 의미 있는 변수가 없을 때 중단하는 방법.
  - **후진제거법(backward elimination method) :** 모든 설명변수를 포함하는 모형에서 기여도가 낮은 변수 를 차례로 제거해 나가되 더 이상 제거할 변수가 없을 때 중단하는 방법.
  - **단계별선택법(stepwise selection method) :** 전진선택법에서 한번 선택된 변수는 다음 단계에서 제거 될 기회를 갖지 못한다(후진제거법의 경우는 반대임). 단계별 선택법은 이러한 단점을 보완한 방법으로 먼저 선택된 변수도 다음 단계에서 제거될 수 있도록 변수선택의 매 단계마다 체크해 나가는 방법.

4.3 변수선택법

- R의 step() 함수는 자동화된 변수선택을 제공한다. 이 함수의 일반 형식은 다음과 같다.

```
step(object, scope, scale = 0, direction = c("both", "backward",
        "forward"), trace = 1, keep = NULL, steps = 1000, k = 2, ...)
```

- object= ”lm” 또는 “glm” 객체
- scope= 고려할 모형의 범위를 지정

  (예) scope=list(lower=null, upper=full), null과 full은 “lm” 객체임
- direction= 변수선택 방법 지정. “both”(단계별선택법)는 디폴트임

4.3 변수선택법

예제 2 swiss 자료를 이용하여 다중선형회귀에서의 여러 가지 변수선택법을 적용한다.

```
> data(swiss)
> str(swiss)
'data.frame':  47 obs. of  6 variables:
 $ Fertility        : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
 $ Agriculture      : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
 $ Examination      : int  15 6 5 12 17 9 16 14 12 16 ...
 $ Education        : int  12 9 5 7 15 7 7 8 7 13 ...
 $ Catholic         : num  9.96 84.84 93.4 33.77 5.16 ...
 $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

자료 설명 이 자료는 1988년 스위스의 출산(측도)과 사회경제지표에 대한 자료이다. Fertility를 반응변수로, 나머지를 예측변수로 하는 다중회귀모형을 적합하고자 한다.

4.3 변수선택법

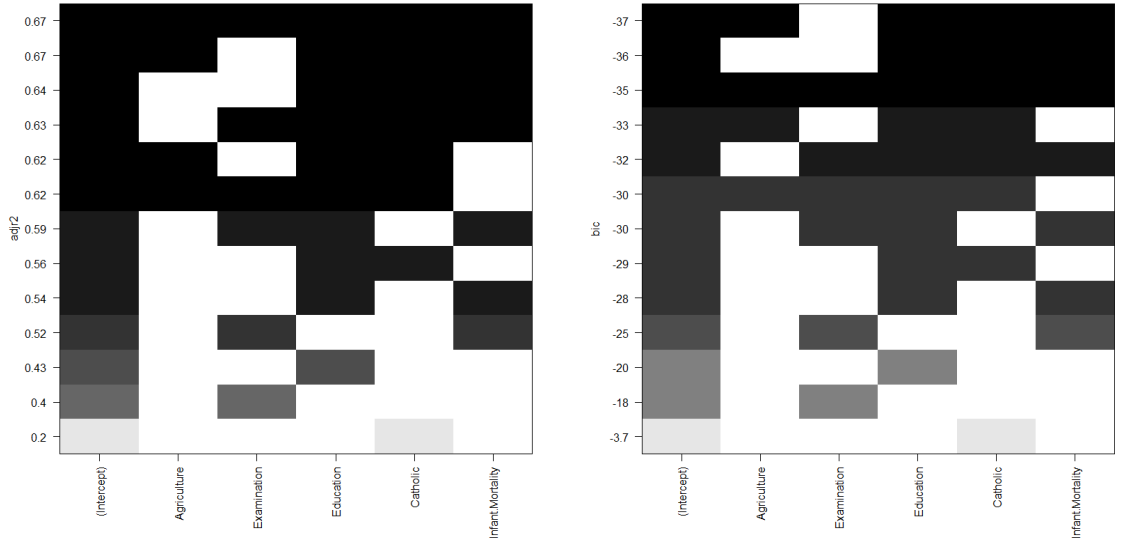
```
> ## (a) 모든 가능한 회귀 적용: regsubsets{leaps} 함수 이용
> library(leaps)
> a <- regsubsets(x=Fertility~.,data=swiss,nbest=3)
> summary(a)
Subset selection object
Call: regsubsets.formula(x = Fertility ~ ., data = swiss, nbest = 3)
5 Variables  (and intercept)
                 Forced in Forced out
Agriculture          FALSE      FALSE
Examination          FALSE      FALSE
Education            FALSE      FALSE
Catholic             FALSE      FALSE
Infant.Mortality     FALSE      FALSE
3 subsets of each size up to 5
```

4.3 변수선택법

```
Selection Algorithm: exhaustive
          Agriculture Examination Education  Catholic Infant.Mortality
1  ( 1 ) " "         " "         "*"        " "      " "
1  ( 2 ) " "         "*"         " "        " "      " "
1  ( 3 ) " "         " "         " "        "*"      " "
2  ( 1 ) " "         " "         "*"        "*"      " "
2  ( 2 ) " "         " "         "*"        " "      "*"
2  ( 3 ) " "         "*"         " "        " "      "*"
3  ( 1 ) " "         " "         "*"        "*"      "*"
3  ( 2 ) "*"         " "         "*"        "*"      " "
3  ( 3 ) " "         "*"         "*"        " "      "*"
4  ( 1 ) "*"         " "         "*"        "*"      "*"
4  ( 2 ) " "         "*"         "*"        "*"      "*"
4  ( 3 ) "*"         "*"         "*"        "*"      " "
5  ( 1 ) "*"         "*"         "*"        "*"      "*"
```

4.3 변수선택법

```
> par(mfrow=c(1,2))
> plot(a, scale="adjr2")
> plot(a, scale="bic")
```



4.3 자동화된 변수선택법

```
> summary(lm1 <- lm(Fertility ~ ., data = swiss))

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213
                                (…)
```

4.3 자동화된 변수선택법

```
                                  (…)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

4.3 자동화된 변수선택법

```
> ## (b) 자동화된 변수선택법: step() 함수 이용
> slm1 <- step(lm1, data = swiss)
> summary(slm1)
Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic +
    Infant.Mortality, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422
                                (…)
```

4.3 변수선택법

```
                                  (…)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

4.3 변수선택법

- 다음의 [예제 3] 은 lm() 함수의 반복 적용을 통해, 수동적인 방법으로, 최적모형을 찾아가는 과정을 보여준다. 자동화된 변수선택법(step() 함수 적용)의 결과도 함께 제시하고 그 결과를 비교한다.
- 아울러 모형에 포함된 예측변수들 간의 상대적인 중요도를 파악하기 위해 표준화된 예측변수를 사용한 결과를 제시한다.

4.3 변수선택법

예제 3 분석에 사용될 자료(state.x77)는 미국의 50개 주에서 여러 변수값(인구, 수입, 문맹비율, 기대수명, 살인율, 고졸비율, 연평균영하기온일수, 면적)을 측정한 자료이다. 이 가운데 기대수명(Life Exp)을 반응변수로 하여 다중회귀분석을 실시한다. 모든 변수는 연속형이다.

```
> data(state)        # state.x77은 행렬 객체임
```

유 의 다중상관계수는 그 의미상 제곱다중상관계수(squared multiple correlation coefficient)의 표현이 더 정확하나, 통상적으로 "제곱"의 표현을 생략하고 사용한다.

```
> st <- as.data.frame(state.x77)
> str(st)

> colnames(st)[4] <- "Life.Exp"    # 변수명에 빈칸을 제외
> colnames(st)[6] <- "HS.Grad"
```

4.3 변수선택법

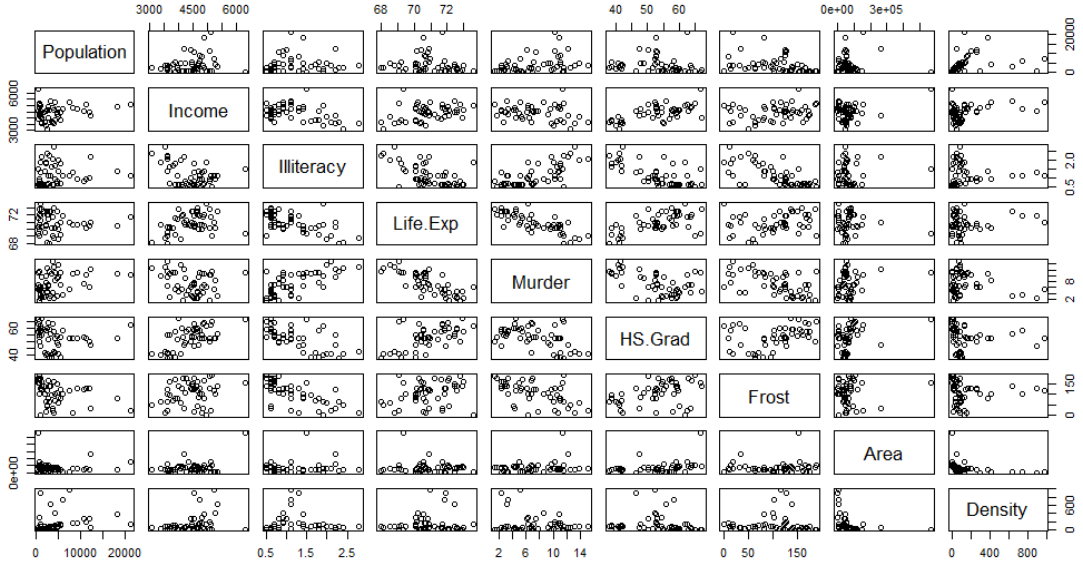
- 인구밀도를 나타내는 새로운 변수(Density)를 생성한다(이를 파생변수(derived variable)라 함). 잘 고안된 파생변수는 회귀모형에서 매우 중요한 변수로 작용할 수 있다.

```
> st[,9] <- st$Population*1000/st$Area
> colnames(st)[9] <- "Density"    # 새로운 열을 생성하고 이름 부여
> str(st)
'data.frame':    50 obs. of  9 variables:
 $ Population: num  3615 365 2212 2110 21198 ...
 $ Income    : num  3624 6315 4530 3378 5114 ...
 $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
 $ Life.Exp  : num  69 69.3 70.5 70.7 71.7 ...
 $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
 $ HS.Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6
 $ Frost     : num  ...
 $ Area      : num  20 152 15 65 20 166 139 103 11 60 ...
 $ Density   : num  50708 566432 113417 51945 156361 ...
                    71.291 0.644 19.503 40.62 135.571 ...
```

4.3 변수선택법

- 자료에 대한 기초분석(기술통계, 상관계수)을 수행한다.

```
> summary(st)
> cor(st)
> pairs(st)
```



4.3 변수선택법

- lm() 또는 step() 함수를 통해 다중회귀모형을 적합한다.
- lm() 함수를 통해 다중회귀모형을 적합하는 과정은 다음과 같다. 단계별로 p-값이 가장 큰 변수를 하나씩 제거해 나가면서 모형을 적합한다(모든 변수의 p-값이 0.05보다 작아질 때 까지).

```
> model1 <- lm(Life.Exp ~ ., data=st)
> summary.aov(model1)
             Df Sum Sq Mean Sq F value   Pr(>F)
Population    1  0.409   0.409   0.760  0.38849
Income        1 11.595  11.595  21.541 3.53e-05
Illiteracy    1 19.421  19.421  36.081 4.23e-07
Murder        1 27.429  27.429  50.959 1.05e-08
HS.Grad       1  4.099   4.099   7.615  0.00861
Frost         1  2.049   2.049   3.806  0.05792
Area          1  0.001   0.001   0.002  0.96438
Density       1  1.229   1.229   2.283  0.13847
Residuals    41 22.068   0.538
```

4.3 변수선택법

```
> model2 <- update(model1, .~.-Area)
> summary(model2)
> anova(model1, model2)   # 모형간 비교
> model3 <- update(model2,  .~.-Illiteracy)
> summary(model3)
> model4 <- update(model3,  .~.-Income)
> summary(model4)
> model5 <- update(model4,  .~.-Density)
> summary(model5)
> model6 <- update(model5,  .~.-Population)
```

4.3 변수선택법

```
> summary(model6)

Call:                                HS.Grad + Frost, data = st)
lm(formula = Life.Exp ~ Murder +

Residuals:                              Max
    Min      1Q  Median      3Q

-1.5015 -0.5391  0.1014  0.5921  1.2268
Coefficients:

(Intercept)  Estimate Std. Error t value  Pr(>|t|)
            71.036379   0.983262  72.246   < 2e-16
Murder      -0.283065   0.036731  -7.706 8.04e-10
HS.Grad      0.049949   0.015201   3.286  0.00195
Frost       -0.006912   0.002447  -2.824  0.00699
                                 (…)
```

4.3 변수선택법

```
                                    (…)
Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

해 석 인구(Population)와 고졸비율(HS.Grad)이 높을수록 기대수명(Life.Exp)은 증가하는 반면, 살인율(Murder)과 연평균영하기온일수(Frost)가 높을수록 기대수명이 줄어듦을 알 수 있다.

4.3 변수선택법

- confint() 함수를 통해 회귀계수에 대한 신뢰구간을 구할 수 있다.

```
> confint(model.step)
                    2.5 %         97.5 %

(Intercept)   6.910798e+01  72.9462729104
Population   -4.543308e-07   0.0001007343
Murder       -3.738840e-01  -0.2264135705
HS.Grad       1.671901e-02   0.0764454870
Frost        -1.081918e-02  -0.0010673977
```

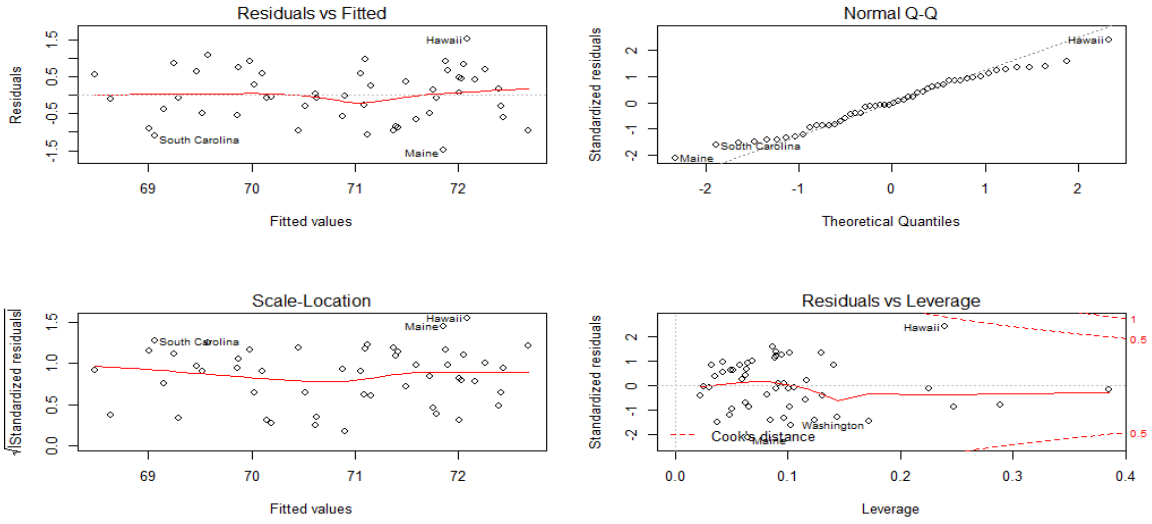
- predict() 함수를 통해 주어진 자료에 대한 예측값을 구할 수 있다.

```
> predict(model.step, list(Population=4000, Murder=10.5,
          HS.Grad=48, Frost=100))

       1

69.71774
```

4.3 변수선택법

- 여러 가지 회귀진단의 결과는 다음과 같다.

```
> par(mfrow=c(2,2))
> plot(model.step)
```



4.3 변수선택법

해 석

그림1(좌측 상단)은 적합값에 대한 잔차 그림으로, 어떤 특별한 패턴을 보이지 않으므로 적합된 선형모형이 적절하다고 할 수 있다. 그림2(우측 상단)는 정규확률그림으로 점들이 비교적 직선 상에 잘 위치하므로 잔차의 정규성 가정이 잘 만족된다고 할 수 있다. 그림3(좌측 하단)은 적합 값에 대한 |표준화 잔차|의 제곱근으로, 정규분포의 가정을 잘 만족하는 것으로 판단된다. 그림 4(우측 하단)는 지렛값에 대한 표준화 잔차와 영향점 진단을 위한 쿡의 거리를 보여준다. 지렛값 이 큰 점이 몇 개 보이며(영향점), 동시에 큰 잔차를 가지는 점(이상치)이 한 개 포함되어 있다.

4.3 변수선택법

- 다중회귀모형의 자세한 적합결과는 다음의 방법으로 추출(확인)할 수 있다.

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

> model.step[[1]]
  (Intercept)    Population        Murder       HS.Grad          Frost
 7.102713e+01  5.013998e-05 -3.001488e-01  4.658225e-02 -5.943290e-03
```

4.3 변수선택법

```
> model.step[[2]]
      Alabama        Alaska       Arizona      Arkansas    California
   0.56888134   -0.54740399   -0.86415671    1.08626119   -0.08564599
     Colorado   Connecticut      Delaware       Florida      Georgia
   0.95645816    0.44541028   -1.06646884    0.04460505   -0.09694227
                                (생략)
  -0.06691392   -0.96272426   -0.96982588    0.47004324   -0.58678863


> sort(model.step$resid)
       Maine South Carolina      Delaware  West Virginia    Washington
  -1.47095411   -1.10109172   -1.06646884   -0.96982588   -0.96272426
Pennsylvania   Mississippi       Arizona       Montana    New Jersey
  -0.95045527   -0.91535384   -0.86415671   -0.84024805   -0.66612086
                                (생략)
North Dakota         Texas      Colorado      Arkansas        Hawaii
   0.90350550    0.92114057    0.95645816    1.08626119    1.50683146
```

4.3 변수선택법

- 베타계수(beta coefficients) 또는 표준화계수(standardized coefficients)는 모든 변수들이 표준화되었을 때의 회귀계수를 의미한다. 베타계수는 예측변수들의 상대적인 중요도를 비교하는 데 유용하다(비표준화 계수 또는 p-값만으로는 상대적인 중요도를 알 수 없다). 모든 변수에 scale() 함수를 적용한 후 lm()을 수행한다.

```
> model.beta <- lm(Life.Exp ~ scale(Population) + scale(Murder) +
                   scale(HS.Grad) + scale(Frost), data=st)

> summary(model.beta)

Call:
lm(formula = Life.Exp ~ scale(Population) + scale(Murder) +
scale(HS.Grad) +
    scale(Frost), data = st)
                                    (…)
```

4.3 변수선택법

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        70.8786     0.1018 696.392  < 2e-16 ***
scale(Population)   0.2238     0.1121   1.996  0.05201 .
scale(Murder)      -1.1080     0.1351  -8.199 1.77e-10 ***
scale(HS.Grad)      0.3762     0.1198   3.142  0.00297 **
scale(Frost)       -0.3089     0.1258  -2.455  0.01802 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

해 석 베타계수의 절대값을 비교해보면 변수의 중요도는 Murder > HS.Grad > Frost > Population 순이다.