4. 로지스틱 회귀모형

4) 변수 선택

4. 변수 선택

- 반응변수에 영향을 줄 수 있다고 생각되는 많은 설명변수 중에서 '최적'의 변수를 선택하여 모형에 포함시키는 절차

- 두 가지 방법으로 구분

1. 검정에 의한 방법
2. 모형선택 기준에 의한 방법

- 어떤 모형을 '최적' 모형으로 정의할 것인가?
- 모수 절약의 원칙

4.1) 검정에 의한 방법

- 검정에 의하여 단계적으로 변수 선택
  - 전진선택법, 후진소거법, 단계적 선택법

- 장점:
  - 계산과정이 비교적 단순
  - 다수의 설명변수가 있는 대규모 자료에 손쉽게 적용 가능

- 단점:

1. 각 단계마다 여러 검정이 동시에 진행
   - 일종오류의 증가
2. 모형 수립 목적이 예측인 경우
   - 변수 선택과정이 목적과 잘 어울리지 않음
3. 변수의 선택과 제거가 '한번에 하나씩' 이루어짐
   - '최적' 모형을 놓치는 경우 발생 가능

- SAS, SPSS
  - 일반적으로 이루어지는 변수 선택방법
  - '최종' 모형을 한 번 실행으로 얻을 수 있음

- R
  - 한 번의 실행으로 '최종' 모형을 얻는 방법은 없음
  - 함수 addterm( ) 혹은 dropterm( )을 반복적으로 실행시켜 사용자가 추가할 변수나 제거할 변수를 직접 선택

    - `addterm(object, scope, test="Chisq")`
      object: 함수 glm( )으로 생성된 객체
      scope: 고려되는 모든 설명변수가 다 포함된 모형
      test="Chisq" : LRT 검정에 의한 변수 추가 결정

    - `dropterm(object, test="Chisq")`

- 전진 선택법

  - 절편만 있는 모형에서 시작하여 영향력이 큰 변수를 각 단계마다 한 개씩 추가하는 방법

    - 영향력이 가장 큰 변수: LRT 검정 통계량의 값이 가장 큰 변수
    - 변수 추가: 영향력이 가장 큰 변수가 유의한 경우

  - 일단 모형에 포함된 변수는 제거 불가

- 예제 4.1: 전진선택법에 의한 모형 선택

```
> library(carData)
> fit_full <- glm(lfp ~ ., family=binomial, Mroz)
> fit <- glm(lfp ~ 1, family=binomial, Mroz)
```

단계 1: 절편만 있는 모형에서 시작. 영향력이 가장 큰 변수 선택.

```
> library(MASS)
> addterm(fit, fit_full, test="Chisq")
```

```
Model:
lfp ~ 1
       Df Deviance     AIC    LRT   Pr(Chi)
<none>     1029.75 1031.75
k5      1   994.75  998.75 34.996 3.303e-09 ***
k618    1  1029.74 1033.74  0.004 0.9469668
age     1  1024.86 1028.86  4.885 0.0270968 *
wc      1  1014.67 1018.67 15.076 0.0001033 ***
hc      1  1027.77 1031.77  1.980 0.1593579
lwg     1  1004.01 1008.01 25.739 3.908e-07 ***
inc     1  1019.31 1023.31 10.438 0.0012347 **
```

변수 k5가 가장 큰 LRT 통계량 값을 갖고 있으며 유의함 → 모형에 포함

단계 2: 변수 k5 모형에 포함시키고 나머지 변수 중 포함 가능 변수 탐색

```
> fit <- update(fit, . ~ . + k5)

> addterm(fit, fit_full, test="Chisq")
Single term additions

Model:
lfp ~ k5
       Df Deviance     AIC     LRT   Pr(Chi)
<none>      994.75  998.75
k618    1   994.53 1000.53  0.2246  0.635523
age     1   964.48  970.48 30.2664 3.766e-08 ***
wc      1   972.98  978.98 21.7728 3.069e-06 ***
hc      1   989.28  995.28  5.4721  0.019322 *
lwg     1   969.17  975.17 25.5782 4.248e-07 ***
inc     1   984.79  990.79  9.9580  0.001602 **
```

변수 age 포함

단계 3: 변수 age 모형에 포함시키고 나머지 변수 중 포함 가능 변수 탐색

```
> fit <- update(fit, . ~ . + age)

> addterm(fit, fit_full, test="Chisq")
Single term additions

Model:
lfp ~ k5 + age
       Df Deviance    AIC     LRT   Pr(Chi)
<none>      964.48 970.48
k618    1   960.71 968.71  3.7762  0.051986 .
wc      1   942.86 950.86 21.6202 3.323e-06 ***
hc      1   960.97 968.97  3.5162  0.060771 .
lwg     1   937.62 945.62 26.8676 2.179e-07 ***
inc     1   956.75 964.75  7.7287  0.005435 **
```

변수 lwg 포함

단계 4: 변수 lwg 모형에 포함시키고 나머지 변수 중 포함 가능 변수 탐색

```
> fit <- update(fit, . ~ . + lwg)

> addterm(fit, fit_full, test="Chisq")
Single term additions

Model:
lfp ~ k5 + age + lwg
       Df Deviance    AIC     LRT  Pr(Chi)
<none>      937.62 945.62
k618    1   935.51 945.51  2.1108 0.146264
wc      1   927.09 937.09 10.5302 0.001174 **
hc      1   936.73 946.73  0.8894 0.345628
inc     1   925.17 935.17 12.4500 0.000418 ***
```

변수 inc 포함

단계 5: 변수 inc 모형에 포함시키고 나머지 변수 중 포함 가능 변수 탐색

```
> fit <- update(fit, . ~ . + inc)

> addterm(fit, fit_full, test="Chisq")
Single term additions

Model:
lfp ~ k5 + age + lwg + inc
       Df Deviance    AIC     LRT   Pr(Chi)
<none>      925.17 935.17
k618    1   923.76 935.76  1.4043   0.23601
wc      1   906.46 918.46 18.7105 1.521e-05 ***
hc      1   919.32 931.32  5.8497   0.01558 *
```

변수 wc 포함

단계 6: 변수 wc 모형에 포함시키고 나머지 변수 중 포함 가능 변수 탐색

```
> fit <- update(fit, . ~ . + wc)

> addterm(fit, fit_full, test="Chisq")
Single term additions

Model:
lfp ~ k5 + age + lwg + inc + wc
       Df Deviance    AIC     LRT Pr(Chi)
<none>      906.46 918.46
k618    1   905.56 919.56 0.89523  0.3441
hc      1   906.17 920.17 0.28675  0.5923
```

변수 k618 비유의적 → 전진선택법 중지

- 후진 소거법

  - 고려 대상이 되는 모든 설명변수가 포함된 모형에서 시작하여 설명력이 미약한 변수를 하나씩 제거해 가는 방법

    - 설명력이 가장 미약한 변수: LRT 통계량 값이 가장 작은 변수
    - 변수 제거: 설명력이 가장 미약한 변수가 비유의적인 경우

  - 일단 모형에서 제거된 변수는 다시 모형에 포함될 수 없음

- 예 4.1: 후진 소거법

단계 1: 모든 설명변수가 포함된 상태에서 시작.
영향력이 가장 미비한 변수 선택

```
> fit <- glm(lfp ~ ., family=binomial, Mroz)

> dropterm(fit, test="Chisq")
Single term deletions

Model:
lfp ~ k5 + k618 + age + wc + hc + lwg + inc
       Df Deviance    AIC    LRT   Pr(Chi)
<none>      905.27 921.27
k5      1   971.75 985.75 66.484 3.527e-16 ***
k618    1   906.17 920.17  0.903  0.342042
age     1   930.86 944.86 25.598 4.204e-07 ***
wc      1   917.99 931.99 12.724  0.000361 ***
hc      1   905.56 919.56  0.294  0.587489
lwg     1   922.27 936.27 17.001 3.736e-05 ***
inc     1   924.77 938.77 19.504 1.004e-05 ***
```

변수 hc 제거

단계 2: 변수 hc 제거하고 나머지 변수 중 제거 가능 변수 탐색

```
> fit <- update(fit, . ~ . - hc)

> dropterm(fit, test="Chisq")
Single term deletions

Model:
lfp ~ k5 + k618 + age + wc + lwg + inc
       Df Deviance    AIC    LRT   Pr(Chi)
<none>      905.56 919.56
k5      1   971.77 983.77 66.211 4.051e-16 ***
k618    1   906.46 918.46  0.895    0.3441
age     1   932.21 944.21 26.652 2.436e-07 ***
wc      1   923.76 935.76 18.201 1.987e-05 ***
lwg     1   922.61 934.61 17.047 3.647e-05 ***
inc     1   925.31 937.31 19.748 8.834e-06 ***
```

변수 k618 제거

단계 3: 변수 k618 제거하고 나머지 변수 중 제거 가능 변수 탐색

```
> fit <- update(fit, . ~ . - k618)

> dropterm(fit, test="Chisq")
Single term deletions

Model:
lfp ~ k5 + age + wc + lwg + inc
       Df Deviance    AIC    LRT   Pr(Chi)
<none>      906.46 918.46
k5      1   971.77 981.77 65.317 6.377e-16 ***
age     1   934.02 944.02 27.565 1.519e-07 ***
wc      1   925.17 935.17 18.711 1.521e-05 ***
lwg     1   924.31 934.31 17.850 2.390e-05 ***
inc     1   927.09 937.09 20.630 5.571e-06 ***
```

모든 변수가 유의적 → 후진 소거법 중지

- 단계적 선택법

  - 전진 선택법:
    - 일단 포함된 변수에 대한 추가적 검정이 없음
    - 포함될 단계에서 유의적이어도 다른 변수가 포함되면 비유의적이 될 수 있음
    - 모형에 포함된 변수에 대한 추가적 검정 필요

  - 단계적 선택법:
    - 전진 선택법과 동일하게 진행.
    - 모형에 변수가 추가되면 모든 변수를 대상으로 후진 소거법 실시하여 제거할 변수 탐색
    - 변수 선택 종료 조건
      1) 전진 선택법으로 추가할 변수가 없는 경우
      2) 이전 단계에서 제거된 변수가 바로 다음 단계에서 선택되는 경우

- R에서 실시 방법

  1) 절편만 있는 모형에서 시작
  2) addterm( )으로 추가할 변수 선택
  3) update( )로 모형에 변수 추가
  4) dropterm( )으로 제거할 변수 탐색
  5) 제거할 변수가 있는 경우 update( )로 제거
  6) 2~5 단계 반복 실시

- 예 4.1: 단계별 선택법

  - 단계적 선택법으로 변수 선택
  - 전진 선택 및 후진 소거 방법과 결과 비교

4.2) 모형 선택 기준에 의한 방법

- 모형 수립 목적
- 주어진 모형이 목적을 얼마나 잘 만족시키는지를 측정할 수 있는 통계량을 변수 선택 기준으로 활용하는 방법

- 일반적인 회귀모형에서 사용할 수 있는 통계량
  - 결정계수, 수정결정계수, MSE, C_p, AIC, BIC, …
- 로지스틱 회귀모형에서 사용할 수 있는 통계량
  - AIC, BIC, …

- 변수 선택 방법
  1. 모든 가능한 회귀
  2. 단계적 선택법

- 모든 가능한 회귀

  - 설명변수의 모든 가능한 조합에 대하여 모형 선택 기준이 되는 통계량 값을 계산하여 최적 모형 선택

  - 대규모 데이터의 경우 시간이 많이 걸리는 방법
    - 설명변수의 수가 p이면 비교할 모형의 수는 2^p

  - 로지스틱 회귀모형의 경우 적용 가능한 R 함수
    - 패키지 bestglm의 함수 bestglm( )

- 로지스틱 회귀모형에 대한 함수 bestglm( )의 일반적인 사용법

```
bestglm(Xy, family=binomial, IC=c("AIC", "BIC"))
```

- Xy: 데이터 프레임. 반응변수는 마지막 열
- IC : 모형 선택 기준 통계량. 디폴트는 BIC

- 예제 4.1: AIC와 BIC에 의한 모든 가능한 회귀

```
> library(bestglm)
> library(dplyr)

> Xy <- select(Mroz, k5:inc, lfp)

> fit_aic <- bestglm(Xy, family=binomial, IC="AIC")
Morgan-Tatar search since family is non-gaussian.

> fit_bic <- bestglm(Xy, family=binomial)
Morgan-Tatar search since family is non-gaussian.
```

- 최종 모형의 개별 회귀계수 추정 및 검정

```
> fit_aic
AIC
BICq equivalent for q in (0.011227459975897 9, 0.94605924015865)
Best Model:
               Estimate  Std. Error   z value     Pr(>|z|)
(Intercept)  2.90192560 0.542897742  5.345253 9.029102e-08
k5          -1.43180362 0.193195960 -7.411147 1.252119e-13
age         -0.05853213 0.011415654 -5.127357 2.938381e-07
wcyes        0.87237347 0.206389444  4.226832 2.370047e-05
lwg          0.61568412 0.150142747  4.100658 4.119762e-05
inc         -0.03367514 0.007800262 -4.317181 1.580346e-05
```

fit_bic와 동일한 결과

- 최종 모형의 유의성 검정

```
> summary(fit_aic)
Fitting algorithm:  AIC-glm
Best Model:
             df  deviance
Null Model 747  906.4554
Full Model 752 1029.7464

        likelihood-ratio test - GLM

data:  H0: Null Model vs. H1: Best Fit AIC-glm
X = 123.29, df = 5, p-value < 2.2e-16
```

- Best subset model list

```
> fit_aic$Subsets
   Intercept    k5  k618   age    wc    hc   lwg   inc logLikelihood       AIC
0       TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE     -514.8732 1029.7464
1       TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE     -497.3750  996.7500
2       TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE     -482.2418  968.4836
3       TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE     -468.8080  943.6160
4       TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE     -462.1527  932.3053
5*      TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE     -453.2277  916.4554
6       TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE     -452.7801  917.5602
7       TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE     -452.6330  919.2659
```

```
> fit_bic$Subsets
   Intercept    k5  k618   age    wc    hc   lwg   inc logLikelihood       BIC
0       TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE     -514.8732 1029.7464
1       TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE     -497.3750 1001.3741
2       TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE     -482.2418  977.7317
3       TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE     -468.8080  957.4882
4       TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE     -462.1527  950.8016
5*      TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE     -453.2277  939.5758
6       TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE     -452.7801  945.3046
7       TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE     -452.6330  951.6344
```

● 단계적 선택법

- 대규모 자료의 경우 모든 가능한 회귀에 의한 탐색은 시간이 지나치게 많이 걸리는 방법이 됨
- 이러한 경우 탐색 범위를 제한하여 단계적으로 AIC 등을 기준으로 선택하는 방법을 사용할 수 있음

- 사용가능 함수: 패키지 MASS의 함수 stepAIC( )

stepAIC(object, scope, k=2,
direction=c("both", "backward", "forward"))

- object: 함수 glm( )으로 생성된 객체
- scope: 모든 설명변수가 포함된 full 모형의 formula.
생략되면 object에 설정된 모형이 full 모형.
- direction: 단계적 탐색의 방향. 디폴트는 "both".
scope가 생략되면 "backward"가 디폴트.
- k: 탐색에 사용되는 IC. k=2는 AIC, k=log(n)은 BIC.

- 예제 4.1: 전진선택에 의한 단계적 모형 탐색

```
> library(MASS)
> fit <- glm(lfp ~ 1, family=binomial, Mroz)
> fit_full <- glm(lfp ~ ., family=binomial, Mroz)
```

`stepAIC(object, scope)`

- object: 탐색 시작 모형 객체
- scope: 모든 설명변수가 포함된 모형의 formula 입력

  → `scope=lfp ~ k5 + k618 + age + wc + hc + lwg + inc`

  변수의 개수가 많아도 모두 일일이 지정해야 함
  모형 fit_full에서 선언된 모형의 formula를 이용하는 것이 더 간편

```
> formula(fit_full)
lfp ~ k5 + k618 + age + wc + hc + lwg + inc
```

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

```
Start:  AIC=1031.75
lfp ~ 1

       Df Deviance     AIC
+ k5    1   994.75  998.75
+ lwg   1  1004.01 1008.01
+ wc    1  1014.67 1018.67
+ inc   1  1019.31 1023.31
+ age   1  1024.86 1028.86
<none>     1029.75 1031.75
+ hc    1  1027.77 1031.77
+ k618  1  1029.74 1033.74

Step:  AIC=998.75
lfp ~ k5

       Df Deviance     AIC
+ age   1   964.48  970.48
+ lwg   1   969.17  975.17
+ wc    1   972.98  978.98
+ inc   1   984.79  990.79
+ hc    1   989.28  995.28
<none>     994.75  998.75
+ k618  1   994.53 1000.53
- k5    1  1029.75 1031.75
```

```
Step:  AIC=970.48
lfp ~ k5 + age

       Df Deviance     AIC
+ lwg   1   937.62  945.62
+ wc    1   942.86  950.86
+ inc   1   956.75  964.75
+ k618  1   960.71  968.71
+ hc    1   960.97  968.97
<none>     964.48  970.48
- age   1   994.75  998.75
- k5    1  1024.86 1028.86

Step:  AIC=945.62
lfp ~ k5 + age + lwg

       Df Deviance     AIC
+ inc   1   925.17  935.17
+ wc    1   927.09  937.09
+ k618  1   935.51  945.51
<none>     937.62  945.62
+ hc    1   936.73  946.73
- lwg   1   964.48  970.48
- age   1   969.17  975.17
- k5    1   998.70 1004.70
```

```
Step:  AIC=935.17
lfp ~ k5 + age + lwg + inc

       Df Deviance    AIC
+ wc    1   906.46 918.46
+ hc    1   919.32 931.32
<none>      925.17 935.17
+ k618  1   923.76 935.76
- inc   1   937.62 945.62
- age   1   954.08 962.08
- lwg   1   956.75 964.75
- k5    1   984.00 992.00

Step:  AIC=918.46
lfp ~ k5 + age + lwg + inc + wc

       Df Deviance    AIC
<none>      906.46 918.46
+ k618  1   905.56 919.56
+ hc    1   906.17 920.17
- lwg   1   924.31 934.31
- wc    1   925.17 935.17
- inc   1   927.09 937.09
- age   1   934.02 944.02
- k5    1   971.77 981.77
```

```
Call:  glm(formula = lfp ~ k5 + age + lwg + inc + wc, family = binomial,
    data = Mroz)

Coefficients:
(Intercept)           k5          age          lwg          inc        wcyes
    2.90193     -1.43180     -0.05853      0.61568     -0.03368      0.87237

Degrees of Freedom: 752 Total (i.e. Null);  747 Residual
Null Deviance:      1030
Residual Deviance: 906.5        AIC: 918.5
```

- 예제 4.1: 후진소거에 의한 단계적 모형 탐색

```
> stepAIC(fit_full)
```

```
Start:  AIC=921.27
lfp ~ k5 + k618 + age + wc + hc + lwg + inc

       Df Deviance    AIC
- hc    1   905.56 919.56
- k618  1   906.17 920.17
<none>      905.27 921.27
- wc    1   917.99 931.99
- lwg   1   922.27 936.27
- inc   1   924.77 938.77
- age   1   930.86 944.86
- k5    1   971.75 985.75

Step:  AIC=919.56
lfp ~ k5 + k618 + age + wc + lwg + inc

       Df Deviance    AIC
- k618  1   906.46 918.46
<none>      905.56 919.56
- lwg   1   922.61 934.61
- wc    1   923.76 935.76
- inc   1   925.31 937.31
- age   1   932.21 944.21
- k5    1   971.77 983.77
```

```
Step:  AIC=918.46
lfp ~ k5 + age + wc + lwg + inc

       Df Deviance    AIC
<none>      906.46 918.46
- lwg   1   924.31 934.31
- wc    1   925.17 935.17
- inc   1   927.09 937.09
- age   1   934.02 944.02
- k5    1   971.77 981.77
```

전진선택에 의한 모형탐색 결과와 동일

- BIC에 의한 단계적 모형 탐색

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

Call:  glm(formula = lfp ~ k5 + age + wc + lwg + inc, family = binomial,
    data = Mroz)

Coefficients:
(Intercept)           k5          age        wcyes          lwg          inc
    2.90193     -1.43180     -0.05853      0.87237      0.61568     -0.03368

Degrees of Freedom: 752 Total (i.e. Null);  747 Residual
Null Deviance:       1030
Residual Deviance: 906.5  AIC: 918.5
```

변수 선택 방법 적용 시 주의점

- 변수선택: 목적을 이루는 수단일 뿐 목적 자체가 아님
  - 선택된 모형이 '최적' 모형을 의미하는 것은 아님

- 회귀진단, 변수변환 등과 분리된 과정이 아니라 서로 연관된 분석과정

- 이상값 등이 발견되어 분석에서 제외되거나 혹은 변수 변환이 이루어진 경우에는 반드시 변수 선택 과정을 다시 실시

- 연습문제

  - 다음 자료에 대한 '최적' 모형을 선택하라.

    1) p2-2.dat
    2) p2-4.dat