4. 로지스틱 회귀모형

1) 모형의 추정 및 해석

내용

1. 모형의 추정 및 해석(2장)
2. 추론(3.1~3.3장)
3. 모형의 적합도 측정(3.4, 3.5장)
4. 다중 로지스틱 회귀모형 구축(4장)

1. Logistic 회귀모형의 추정 및 해석

- Logistic 회귀모형

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- 모수 $\beta_0, \beta_1, \cdots, \beta_p$ 추정: Maximum Likelihood Estimation
  - 정규분포의 경우와는 다르게 정확한 $\hat{\beta}$을 구할 수 없음
  - 비선형 정규방정식 → 반복 계산에 의한 추정

- 모수 추정에 실패하는 경우
  - 설정된 모형이 적절하다면 몇 번의 반복만으로도 모수 추정 가능

  - 반복 계산 수렴 기준을 충족시키지 못해 추정에 실패하는 경우 발생 가능

1. 관측값의 크기가 충분히 크지 않았을 때
2. 독립변수의 측정 척도가 매우 다를 때
3. 성공 혹은 실패 중 한 범주의 발생 빈도가 매우 낮을 때

Logistic 회귀모형 추정을 위한 R 함수

- GLM을 위한 R 함수: glm( )

- 이항 반응변수인 경우 함수 glm( )의 일반적인 사용법

```
glm(formula, family=binomial, data, ...)
```

  - formula: response ~ terms 형식의 R 공식
    - response: 숫자형 벡터 혹은
      요인(첫 번째 범주가 '실패'
      두 번째 범주가 '성공'으로 처리됨)

  - family: 반응변수의 분포 및 link function
    - 이항 반응변수: binomial
    - link function: 디폴트는 logit(생략됨)
    - probit을 원하는 경우: family=binomial(link="probit")

- 예제 2.2: 부인 직업 참여 여부 결정에 대한 로지스틱 회귀모형 분석

1) logistic 회귀계수 추정

```
> library(carData)
> with(Mroz, table(lfp))
lfp
 no yes
325 428
```

'no': 첫 번째 범주. '실패'로 인식
'yes': 두 번째 범주. '성공'으로 인식

→ 함수 glm( ):
'성공' 확률 P(lfp='yes') 추정

```
> fit1 <- glm(lfp~. , family=binomial, Mroz)
```

- 추정 결과 확인

```
> summary(fit1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1062  -1.0900   0.5978   0.9709   2.1893

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
---

Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.27  on 745  degrees of freedom
AIC: 921.27

Number of Fisher Scoring iterations: 4
```

- p-값을 정규분포에서 계산
- 2장에서는 카이제곱 분포에서 계산
- $Z \sim N(0,1),\ Z^2 \sim \chi^2$
- 변수 wcyes와 hcyes는 가변수

- 모형에서 비유의적인 변수(k618, hc) 제거

```
> fit2 <- glm(lfp~.-k618-hc, family=binomial, Mroz)
```

```
> fit2 <- update(fit1, .~.-k618-hc)
```

```
> summary(fit2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.90193    0.54290   5.345 9.03e-08 ***
k5          -1.43180    0.19320  -7.411 1.25e-13 ***
age         -0.05853    0.01142  -5.127 2.94e-07 ***
wcyes        0.87237    0.20639   4.227 2.37e-05 ***
lwg          0.61568    0.15014   4.101 4.12e-05 ***
inc         -0.03367    0.00780  -4.317 1.58e-05 ***

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  906.46  on 747  degrees of freedom
AIC: 918.46

Number of Fisher Scoring iterations: 3
```

- 추정된 로지스틱 회귀곡선

  - 모든 설명변수 포함

$$\hat{\pi}(\boldsymbol{x}) = \frac{exp(3.18 - 1.46\mathrm{k5} - \cdots - 0.03\mathrm{inc})}{1 + exp(3.18 - 1.46\mathrm{k5} - \cdots - 0.03\mathrm{inc})}$$

  - 비유의적 설명변수(k618, hc) 제외

$$\hat{\pi}(\boldsymbol{x}) = \frac{exp(2.9 - 1.43\mathrm{k5} - \cdots - 0.03\mathrm{inc})}{1 + exp(2.9 - 1.43\mathrm{k5} - \cdots - 0.03\mathrm{inc})}$$

2) 직업참여 확률, $P(\text{lfp} = "yes")$의 추정

- 함수 predict( )에 의한 확률 추정

```
predict(object,  newdata= , type="response")
```

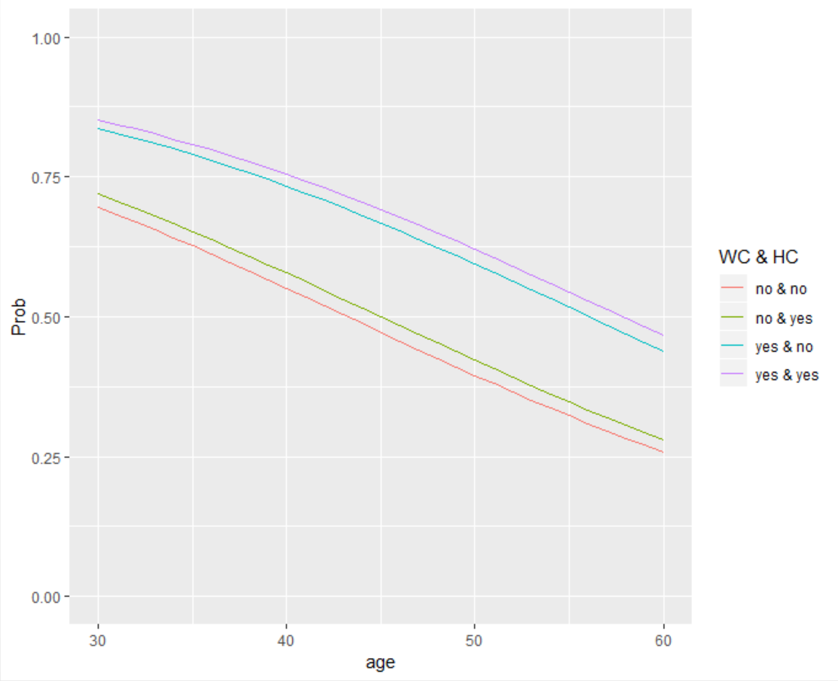
- object: 함수 glm( )으로 생성된 객체

- newdata= : 새로운 설명변수 값으로 구성된 데이터 프레임.
  생략 시 기존 자료에 대한 확률 추정

- type="response" : 반응변수의 scale로 추정
  → $P(\text{lfp} = "yes")$의 추정

- 새로운 설명변수 값에 대한 직업 참여 확률 추정

(1) k5, k618, lwg, inc: 평균값　　age: 30~60　　wc, hc: 4가지 조합



교재 그림 2.3

R code

```
> library(dplyr)
> df1 <- summarize(Mroz, k5=mean(k5), k618=mean(k618),
                         lwg=mean(lwg), inc=mean(inc))
> df1 <- cbind(df1, age=30:60)
```

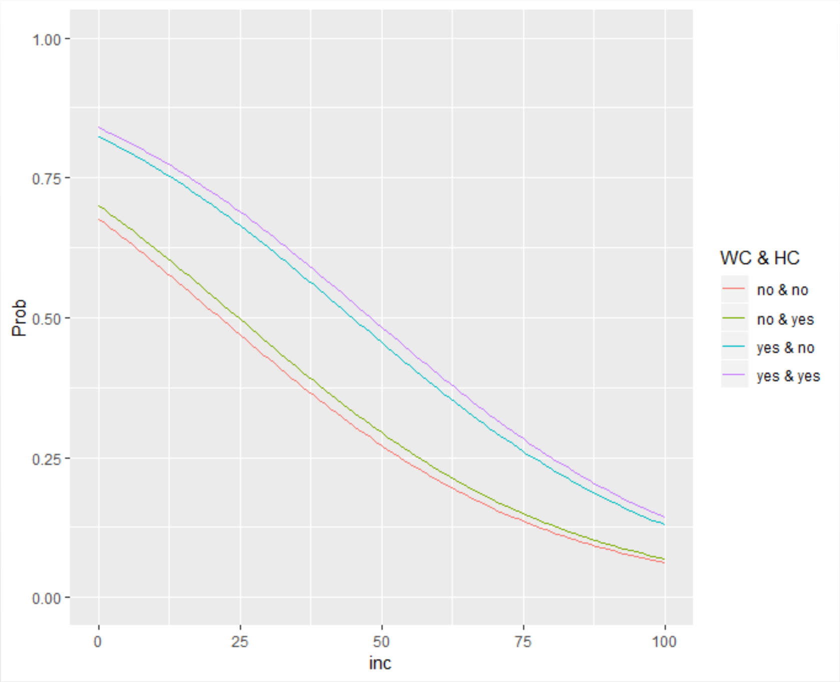
```
> levels(Mroz$wc)
[1] "no"  "yes"
```

```
> prob_1 <- predict(fit1, newdata=cbind(df1, wc="no", hc="no"),
                    type='response')
> prob_2 <- predict(fit1, newdata=cbind(df1, wc="no", hc="yes"),
                    type='response')
> prob_3 <- predict(fit1, newdata=cbind(df1, wc="yes", hc="no"),
                    type='response')
> prob_4 <- predict(fit1, newdata=cbind(df1, wc="yes", hc="yes"),
                    type='response')

> df_2 <- tibble(age=30:60, p1=prob_1, p2=prob_2,
                 p3=prob_3, p4=prob_4)
```

```
> library(ggplot2)
> ggplot(df_2) +
    geom_line(aes(x=age, y=p1, col="no & no")) +
    geom_line(aes(x=age, y=p2, col="no & yes")) +
    geom_line(aes(x=age, y=p3, col="yes & no")) +
    geom_line(aes(x=age, y=p4, col="yes & yes")) +
    ylim(0,1) + labs(y="Prob", col="WC & HC")
```

(2) k5, k618, age, lwg: 평균값 inc: 0~100 wc, hc: 4가지 조합



교재 그림 2.4

R code

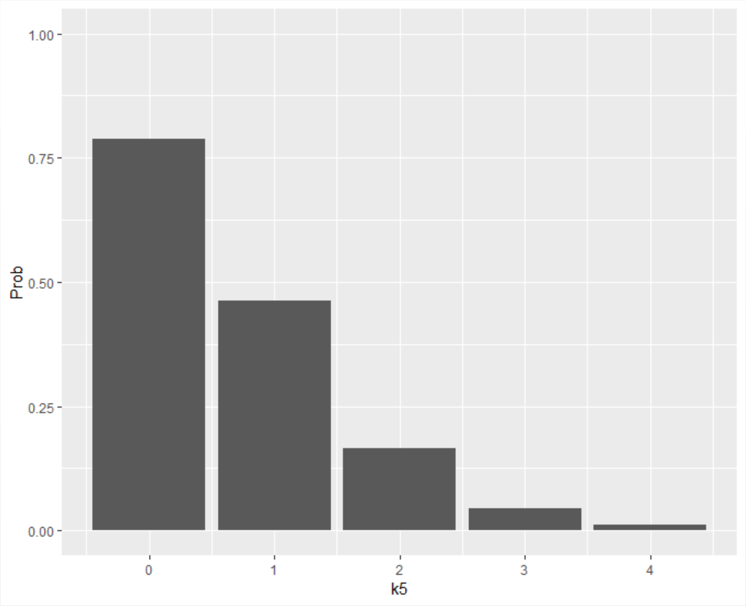
```
> library(dplyr)
> df3 <- summarize(Mroz, k5=mean(k5), k618=mean(k618), age=mean(age),
                   lwg=mean(lwg))
> df3 <- cbind(df3, inc=0:100)

> prob_1 <- predict(fit1, newdata=cbind(df3, wc="no", hc="no"),
                    type='response')
> prob_2 <- predict(fit1, newdata=cbind(df3, wc="no", hc="yes"),
                    type='response')
> prob_3 <- predict(fit1, newdata=cbind(df3, wc="yes", hc="no"),
                    type='response')
> prob_4 <- predict(fit1, newdata=cbind(df3, wc="yes", hc="yes"),
                    type='response')

> df_4 <- tibble(inc=0:100, p1=prob_1, p2=prob_2,
                 p3=prob_3, p4=prob_4)
```

```
> library(ggplot2)
> ggplot(df_4) +
    geom_line(aes(x=inc, y=p1, col="no & no")) +
    geom_line(aes(x=inc, y=p2, col="no & yes")) +
    geom_line(aes(x=inc, y=p3, col="yes & no")) +
    geom_line(aes(x=inc, y=p4, col="yes & yes")) +
    ylim(0,1) + labs(y="Prob", col="WC & HC")
```

(3) k5=0~4 k618, age, lwg, inc: 평균값 wc="yes", hc="yes"



교재 그림 2.5

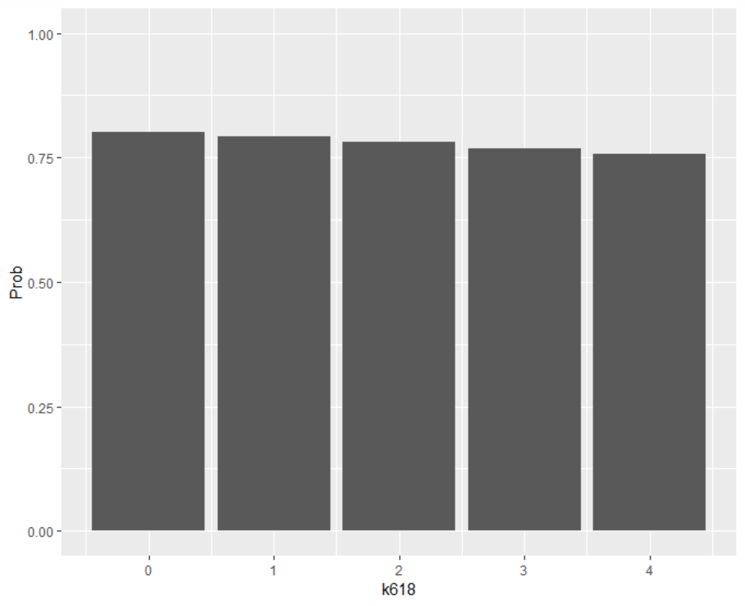
R code

```
> df5 <- summarize(Mroz, k618=mean(k618), age=mean(age),
                   lwg=mean(lwg), inc=mean(inc))
> df5 <- cbind(df5, k5=0:4, wc="yes",hc="yes")

> prob_1 <- predict(fit1, newdata=df5, type="response")
> df_6 <- tibble(k5=0:4, p1=prob_1)

> ggplot(df_6) +
    geom_bar(aes(x=k5, y=p1), stat="identity") +
    labs(y="Prob") + ylim(0,1)
```

(4) k5=0 k618=0~4 age, lwg, inc: 평균값 wc=“yes”, hc=“yes”



교재 그림 2.6

R code

```
> df7 <- summarize(Mroz, age=mean(age), lwg=mean(lwg),
                   inc=mean(inc))
> df7 <- cbind(df7, k5=0, k618=0:4, wc="yes",hc="yes")

> prob_1 <- predict(fit1, newdata=df7, type="response")
> df_8 <- tibble(k618=0:4, p1=prob_1)

> ggplot(df_8) +
    geom_bar(aes(x=k618, y=p1), stat="identity") +
    labs(y="Prob") + ylim(0,1)
```

2. 설명변수의 효과분석

- 선형회귀모형:
  - 다른 설명변수들의 수준을 고정시킨 상태에서 X_j를 한 단위 증가시키면 $E(Y)$는 β_j 만큼 변화

- 로지스틱 회귀모형:
  - 비선형 모형이기 때문에 선형회귀모형의 방식으로 효과분석 불가능
  - 대안

1. 확률의 부분변화 (교재 2.4.1 생략)
2. 확률의 이산변화 (교재 2.4.2 생략)
3. Odds ratios (교재 2.4.3)

- Odds ratio에 의한 설명변수 효과 분석

  - 로지스틱 회귀모형: log(odds)에 대한 모형

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \log\Omega(\boldsymbol{x}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - Odds에 대한 모형

$$\Omega(\boldsymbol{x}) = exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) = e^{\beta_0} e^{\beta_1 X_1} \ldots e^{\beta_j X_j} \ldots e^{\beta_p X_p}$$

  - 설명변수 X_j의 수준을 δ 만큼 변화시켰을 때 odds

$$\Omega(\boldsymbol{x}, X_j + \delta) = e^{\beta_0} e^{\beta_1 X_1} \ldots e^{\beta_j (X_j + \delta)} \ldots e^{\beta_p X_p}$$

  - 설명변수 X_j의 수준을 δ 만큼 변화시켰을 때 odds의 변화: Odds ratio

$\Omega(\boldsymbol{x}, X_j + \delta)/\Omega(\boldsymbol{x}) = e^{\beta_j \delta}$ 변수 X_j의 효과

- 예제 2.2: Odds ratio에 의한 설명변수의 효과분석

1) 로지스틱 회귀모형 및 회귀계수

```
> fit1 <- glm(lfp~. , family=binomial, data=Mroz)
> coef(fit1)
(Intercept)          k5        k618         age       wcyes       hcyes 
   3.182140   -1.462913   -0.064570   -0.062870    0.807273    0.111733 
        lwg         inc 
   0.604693   -0.034446 
```

log(Odds)에 대한 모형: 지수 변환으로 Odds에 대한 모형으로 변환

2) 각 설명변수의 Odds ratio 계산

```
> exp(coef(fit1))
(Intercept)          k5        k618         age       wcyes       hcyes 
   24.09827     0.23156     0.93746     0.93906     2.24178     1.11821 
        lwg         inc 
    1.83069     0.96614 
```

- Odds ratio에 대한 대략적인 해석
  - 공통 가정: 다른 설명변수의 수준은 고정
  - 1보다 작은 값: 해당 설명변수를 1단위 증가 시켰을 때 부인이 직업을 가질 odds 감소
  - 1보다 큰 값: 해당 설명변수를 1단위 증가 시켰을 때 부인이 직업을 가질 odd 증가

odds의 증감은 확률의 증감을 의미

- 각 설명변수 odds ratio 값에 대한 구체적인 해석

```
> exp(coef(fit1))
(Intercept)          k5        k618         age       wcyes       hcyes 
   24.09827     0.23156     0.93746     0.93906     2.24178     1.11821 
        lwg         inc 
    1.83069     0.96614 
```

다른 설명변수의 수준을 고정시켰을 때

k5를 한 단위 증가시키면 직업에 참여할 odds ratio는
$exp(\hat{\beta}_1) = exp(-1.4629) = 0.232$배 감소
→ 100×(0.232-1)=-76.8 즉 76.8% 감소

k5를 두 단위 증가시키면 직업에 참여할 odds ratio는
$exp(\hat{\beta}_1 \times 2\ ) = exp(-1.4629 \times 2) = 0.0536$배 감소
→ 100×(0.0536-1)=-94.6 즉 94.6% 감소

```
> exp(coef(fit1))
(Intercept)          k5        k618         age       wcyes       hcyes 
   24.09827     0.23156     0.93746     0.93906     2.24178     1.11821 
        lwg         inc 
    1.83069     0.96614 
```

다른 설명변수의 수준을 고정시켰을 때

lwg를 한 단위 증가시키면 직업에 참여할 odds ratio는
$exp(\hat{\beta}_6) = exp(0.6047) = 1.831$배 증가
→ 100×(1.831-1)=83.1 즉 83.1% 증가

lwg를 두 단위 증가시키면 직업에 참여할 odds ratio는
$exp(\hat{\beta}_6 \times 2) = exp(0.6047 \times 2) = 3.35$배 증가
→ 100×(3.35-1)=235.1 즉 235.1% 증가

부인 학력수준(wc)이 대졸인 경우가 고졸 이하의 경우와 비교하여 직업에 참여할 odds ratio는 2.242배 증가
→ 100×(2.242-1)=124.2 즉 124.2% 증가

3) 각 설명변수 odds ratio에 대한 95% 신뢰구간

```
> exp(confint(fit1))
Waiting for profiling to be
done...
                2.5 %     97.5 %
(Intercept) 6.9377228 87.0347916
k5          0.1555331  0.3370675
k618        0.8200446  1.0710837
age         0.9154832  0.9625829
wcyes       1.4347543  3.5387571
hcyes       0.7467654  1.6766380
lwg         1.3689201  2.4768235
inc         0.9502809  0.9814042
```

신뢰구간에 1이 포함되어 있는 변수
- 비유의적 변수
- summary(fit1) 결과와 비교

profile likelihood 방식에 의한 신뢰구간 계산:
- Wald 검정 방식에 의한 교재 표 2.12의 결과와는 약간 다름
- 신뢰구간이 odds ratio 점추정값에 대하여 좌우대칭이 아님

'가설 검정'에서 신뢰구간에 대한 추가 설명 예정

3. Probit 모형

- Probit 모형: $\Phi^{-1}[P(Y=1)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

함수 $\Phi(x)$는 누적 표준정규분포

- 추정된 probit 모형:

$$P(Y=1) = \Phi\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)$$

예제 2.3: 직업 참여자료에 대한 Probit 모형 적합

```
> fit.p <- glm(lfp~. , family=binomial(link="probit"), data=Mroz)

> summary(fit.p)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.918418   0.382356   5.017 5.24e-07 ***
k5          -0.874712   0.114423  -7.645 2.10e-14 ***
k618        -0.038595   0.040950  -0.942 0.345942
age         -0.037824   0.007605  -4.973 6.58e-07 ***
wcyes        0.488310   0.136731   3.571 0.000355 ***
hcyes        0.057172   0.124207   0.460 0.645306
lwg          0.365635   0.089992   4.063 4.85e-05 ***
inc         -0.020525   0.004852  -4.230 2.34e-05 ***
---

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.39  on 745  degrees of freedom
AIC: 921.39

Number of Fisher Scoring iterations: 4
```

- 회귀모형의 회귀계수 추정값 비교(logit vs probit)

```
> cbind(logit=round(coef(fit1),3), probit=round(coef(fit.p),3))
             logit probit
(Intercept)  3.182  1.918
k5          -1.463 -0.875
k618        -0.065 -0.039
age         -0.063 -0.038
wcyes        0.807  0.488
hcyes        0.112  0.057
lwg          0.605  0.366
inc         -0.034 -0.021
```

- 기존 자료에 대한 직업 참여 확률 추정 비교(logit vs probit)

```
> cbind(logit=fit1$fitted, probit=fit.p$fitted)[1:10,]
       logit    probit
1  0.5158291 0.5206967
2  0.6668165 0.6650898
3  0.4565831 0.4643790
4  0.6620169 0.6593693
5  0.6632299 0.6653360
6  0.5959744 0.5958797
7  0.9242061 0.9354251
8  0.6586118 0.6573715
9  0.4738387 0.4785964
10 0.7483850 0.7471961
```

- 거의 동일한 결과
- 회귀계수의 차이는 모형의 다름으로 인한 것

- Probit 모형의 단점: 개별 설명변수의 효과분석에서 로지스틱 회귀모형과는 다르게 odds ratio에 의한 분석 불가능
- 상당한 불편함 초래

4. 적용 분야

- 로지스틱 회귀분석의 주요 목적: 판별분석과 거의 동일

1. 반응변수의 구분을 설명할 수 있는 모형 추정: 두 가지 명목형 범주의 차이를 설명할 수 있는 비선형 모형 추정
2. 각 범주에 속할 확률 추정: 추정된 모형을 근거로 주어진 설명변수 수준에서 각 범주에 속할 확률 추정
3. 범주에 대한 분류: 추정된 확률을 근거로 각 관찰값의 범주를 예측

- 적용 예

1. 중소기업 부실 여부 예측
2. 신상품 구매의사 성향 예측
3. 특정 질환 판정 예측
4. 보험 부당 청구 탐지