범주형 자료분석 개론

2021210088 허지혜

5.5 대체 연결함수 : 선형확률모형과 프로빗 모형

로지스틱 회귀 모형은 이산반응변수에 대해 가장 널리 사용되는 모형.
그렇지만, 다른 연결 함수를 사용하는 모형이 때때로 적절히 해석하기 간단할 수 있다.
다른 연결함수인 선형확률모형과 프로빗 모형을 알아보자.

5.5.1 대체 연결함수 : 선형확률모형

앞에서 선형확률모형에 대해 소개하였다.다중 설명변수들이 여러개 있는 경우 선형확률모형은 아래와 같다.

$$P(Y=1)=\alpha+\beta_1 x_1+\ldots+\beta_p x_p$$

이 모형에서 각 x_j에 대해서 성공확률이 선형적으로 변한다고 가정한다.
이 모형은 이항랜덤성분과 항등연결함수를 가정하는 GLM이다.
이 모형 구조상 예측값은 전체 실수선 상에서 그 값을 가질 수 있지만, 확률은 0과 1사이의 값을 가져야 한다.
이런 제약 때문에 이 모형의 적용범위는 한정적이다.

이항 분포 가정 하에서 ML 추정값을 찾기 위한 반복 알고리즘은 일부 관측값에 대하여 추정 확률값이
0과 1 사이의 범위를 벗어날 때 수렴에 실패한다.
그러면 소프트웨어는 무시하고 수렴과 부족 과 같은 오류를 출력한다.
Y가 이항변수라는 특성을 무시하고 일반적인 회귀모형을 사용하게 되면 확률값은 최소제곱추정값이 된다.
이 추정값은 Y가 일정한 분산을 갖는 정규분포를 따른다는 가정하에서 ML 추정값이 된다.

Y가 정규확률변수의 경우 Y의 평균 추정값은 임의의 실수값이 될 수 있고 0과 1 사이의 범위로 제한되지 않기 때문에
이 ML 추정값은 존재한다.이항분포 가정 하에서 ML 적합이 실패할 경우, 그룹화되지 않은 자료에 최소제곱법을 적용
하여 그 추정값을 성공적으로 구할 수 있지만 일부 추정값은 0과 1 사이의 범위를 벗어날 수 있다.

선형확률모형의 장점 = 추정된 효과를 확률척도에 대해 기울기로 쉽게 해석할 수 있다.
이 효과는 적합된 로지스틱 모형의 적합 결과에 있는 평균 주변효과값과 유사한 추정된 값을 가진다.

5.5.2 예제 : 정치성향과 진화론

y = 진화에 대한 의견(1 = 맞다, 0 = 틀리다)
x = 정치 성향(1 = 매우보수적, … ,7 = 매우진보적)
선형확률모형의 경우 x를 양적으로 취급하여 ML 적합을 구하면 다음과 같은 식이 된다.

$$\hat{P}(Y=1) = 0.108 + 0.110x$$

```
> fit <- glm(evolved~ideology, family=quasi(link=identity,
+ variance="mu(1-mu)"), data=Evo)
>
> summary(fit, dispersion=1)

Call:
glm(formula = evolved ~ ideology, family = quasi(link = identity,
    variance = "mu(1-mu)"), data = Evo)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0558  -1.2617   0.7247   1.0954   1.7440

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.108437   0.038932   2.785  0.00535 **
ideology    0.110102   0.008971  12.273  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 1)

    Null deviance: 1469.3  on 1063  degrees of freedom
Residual deviance: 1359.5  on 1062  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
```

```
> fit2 <-glm(evolved~ideology, family=gaussian(link=identity),
+  data=Evo)
> summary(fit2)

Call:
glm(formula = evolved ~ ideology, family = gaussian(link = identity),
    data = Evo)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8819  -0.5492   0.2290   0.4508   0.7836

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10554    0.04272    2.47   0.0137 *
ideology     0.11091    0.01033   10.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2247506)

    Null deviance: 264.57  on 1063  degrees of freedom
Residual deviance: 238.69  on 1062  degrees of freedom
AIC: 1435.2

Number of Fisher Scoring iterations: 2
```

5.5.2 예제 : 정치성향과 진화론

```
> fit <- glm(evolved~ideology, family=quasi(link=identity,
+ variance="mu(1-mu)"), data=Evo)
>
> summary(fit, dispersion=1)

Call:
glm(formula = evolved ~ ideology, family = quasi(link = identity,
    variance = "mu(1-mu)"), data = Evo)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0558  -1.2617   0.7247   1.0954   1.7440

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.108437   0.038932   2.785  0.00535 **
ideology    0.110102   0.008971  12.273  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 1)

    Null deviance: 1469.3  on 1063  degrees of freedom
Residual deviance: 1359.5  on 1062  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
```

```
> fit2 <-glm(evolved~ideology, family=gaussian(link=identity),
+  data=Evo)
> summary(fit2)

Call:
glm(formula = evolved ~ ideology, family = gaussian(link = identity),
    data = Evo)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8819  -0.5492   0.2290   0.4508   0.7836

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10554    0.04272    2.47   0.0137 *
ideology     0.11091    0.01033   10.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2247506)

    Null deviance: 264.57  on 1063  degrees of freedom
Residual deviance: 238.69  on 1062  degrees of freedom
AIC: 1435.2

Number of Fisher Scoring iterations: 2
```

진화를 믿는 확률의 추정값이 정치성향이 진보적인 방향으로 한 범주씩 바뀔 때마다 0.11씩 증가한다.
이 추정 효과는 해석하기 간단하고 유용한 값이다.
이에 대응되는 로지스틱 모형의 적합 결과는 $\text{logit}[\hat{P}(Y=1)] = -1.757 + 0.494x$ 이 되고
평균주변효과에 대한 평균값은 0.111이다.

선형확률모형으로부터 구한 진화를 믿을 확률의 추정값은 매우 보수적인 경우의 0.218 값과
매우 진보적인 경우의 0.879값 사이에 있다.
이 추정값은 이 자료에 대한 그룹화된 분석으로부터 구한 이탈도 통계량값 3.45(df=5)에서
보듯이 각 범주별로 구한 표본비율값을 잘 설명한다.

5.5.3 프로빗 모형과 정규 잠재변수모형

로지스틱 회귀모형은 다음과 같은 S자 모양의 곡선을 가진 또 다른 모형을 프로빗 모형이라고 부른다.
프로빗 모형의 식은 다음과 같다.

$$\text{probit}[P(Y=1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

프로빗 연결함수라고 불리는 모형의 연결함수는 P(Y=1)를 표준정규분포의 왼쪽 꼬리확률이 P(Y=1)와 동일한 표준정규분포 Z 점수로 변환한다.

프로빗 모형은 일반 선형모형과 관련시킬 수 있을 때 프로빗 모형 모수의 해석을 가장 간단하게 할 수 있다.
많은 이항변수는 관측할 수 없는 연속 변수값에 대한 대략적인 측정값으로 간주할 수 있다.

예로 y = 정치성향에 대해 이항 회귀모형을 사용한다고 가정할 때,
여기서 연구에 참여하는 각 사람들은 본인의 정치성향이 진보인지 보수인지 선택해야 한다.
실질적으로 같은 정치성향 범주에 속하는 사람들 사이에서도 정치성향의 차이가 존재한다.
정치성향을 정교하게 측정할 수 있는 방법을 사용하면 진보와 보수 사이에 정치성향이 본질적으로
연속적인 측도라는 것을 상상해볼 수 있다.
즉, 각 사람들은 매우 진보적인 정치성향에서 시작해서 매우 보수적인 성향을 가질 수 있으며
또 이 사이에 아주 많은 가능한 정치성향이 존재할 수 있다.

통계학에서 실제로 관측한 자료에 잠재해 있는 관측되지 않은 변수를 잠재변수라고 부른다.

5.5.3 프로빗 모형과 정규 잠재변수모형

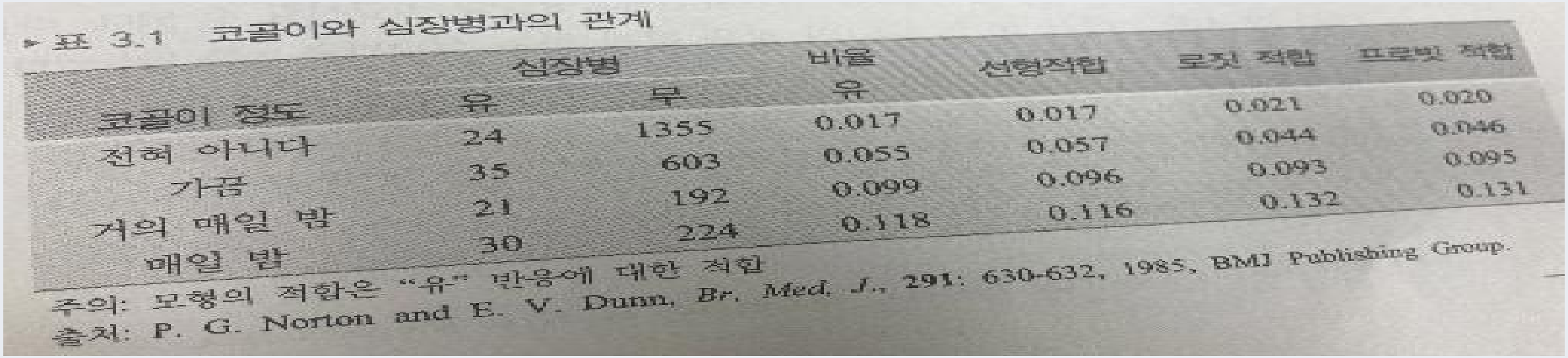
어떤 알려지지 않은 분계점 τ 에 대해 잠재변수 y^* 가 있다고 가정하자.
만약 $y^* \leq \tau$ 일 때 y=0이 관측되고 반대일때 y=1이 관측된다고 하자.
잠재변수 y^* 가 일반적인 선형모형을 만족한다고 가정하자.

$$y^* = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

오차항 ϵ 이 설명변수의 모든 값에 대해서 동일한 분산을 갖는 정규분포를 가지는 경우, 관측된 이항반응변수 y 에 대해서는 프로빗 모형이 만족이 되어야 한다.
더 나아가 var(ϵ)=1이 되도록 잠재변수를 척도변환을 하면 프로빗 모형의 효과는 잠재변수모형의 효과와 동일하다.

그러면 프로빗 모형의 $\hat{\beta}_j$ 를 다른 설명변수값을 보정한 후에 x_j의 한 단위 증가에 따른
E(y^*)의 변화에 대한 추정값으로 해석할 수 있다.
var(τ)에 대한 임의의 값을 갖는 경우에는 $\hat{\beta}_j$ 는 y^* 분포가 표준편차의 몇 배만큼 이동하였는지를 나타낸다.

5.5.4 예제 : 코골이와 심장병 자료의 재분석



▸표 3.1 코골이와 심장병과의 관계

코골이 정도	심장병 유	심장병 무	비율 유	선형적합	로짓 적합	프로빗 적합
전혀 아니다	24	1355	0.017	0.017	0.021	0.020
가끔	35	603	0.055	0.057	0.044	0.046
거의 매일 밤	21	192	0.099	0.096	0.093	0.095
매일 밤	30	224	0.118	0.116	0.132	0.131

주의: 모형의 적합은 "유" 반응에 대한 적합

출처: P. G. Norton and E. V. Dunn, *Br. Med. J.*, **291**: 630-632, 1985, BMJ Publishing Group.

사진은 코골이가 심장병에 미치는 잠재적 영향을 보기 위한 연구 자료이다.
실제 반응변수는 이항변수이지만 심장병 정도를 나타내는 잠재적인 연속형 측도 y^*를 생각해볼 수 있다.
코골이 정도에 대한 점수(0,2,4,5)를 가장한 프로빗 모형의 ML적합 결과는 다음과 같다.

y^*에 대한 잠재변수 모형의 경우, x = 코골이 정도가 한 단위 증가하면 E(y^*)가 표준 편차의 0.188배 만큼 증가하게 된다. x가 0에서 5로 증가함에 따라 심장병의 잠재분포는 거의 표준편차만큼 오른쪽으로 이동한다.

코골이 정도가 x=0일 경우, probit은 -2.601+0.188(0) = -2.06이다.
따라서 심장병을 가질 확률은 표준정규분포에서 z=-2.06의 왼쪽 꼬리확률값이 되고 그 값은 0.002이다.
코골이 정도가 x=5일 경우, probit은 -2.061+0.188(5) = -1.12이며 대응되는 확률값은 0.131이다.

5.5.4 예제 : 코골이와 심장병 자료의 재분석

```
> Heart <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Heart.dat",
+ header=TRUE)
> Heart
             snoring yes   no
1              never  24 1355
2         occasional  35  603
3 nearly_every_night  21  192
4        every_night  30  224
> library(dplyr)
> Heart$x <- recode(Heart$snoring, never=0, occasional=2,
+ nearly_every_night=4, every_night=5)
오류가 있음:
Unreplaced values treated as NA as .x is not compatible. Please specify replacements exhaustively or supply .default
>
> fit <- glm(yes/(yes+no)~x, family=binomial(link=probit),
+ weights=yes+no, data=Heart)
> summary(fit)

Call:
glm(formula = yes/(yes + no) ~ x, family = binomial(link = probit),
    data = Heart, weights = yes + no)

Deviance Residuals:
      1       2       3
-0.3949  0.7052  -0.6940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.08262    0.07435 -28.010  < 2e-16 ***
x            0.81307    0.03401  23.903 4.33e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 39.08575  on 3  degrees of freedom
Residual deviance:  0.85961  on 1  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 19.993

Number of Fisher Scoring iterations: 3

> fitted(fit)
          1           2           3
0.016642998 0.048732834 0.165980587
```

5.5.5 잠재변수모형과 이항회귀모형의 함축적 관계

오차항에 대해 다른 분포를 가정하게 되면 잠재변수의 구조가 달라져서 프로빗 모형과는 다른 이항회귀모형을 얻게 된다. 예로, 로지스틱 분포는 정규분포와 비슷하지만 약간 더 두꺼운 꼬리를 가지고 있다.

오차항에 대해 로지스틱 분포를 갖는 잠재변수 모형은 관찰된 이항반응변수에 대해 로짓 연결함수와 로지스틱 회귀모형을 함축한다.

실제로 정규분포와 로지스틱 분포가 매우 유사하기 때문에 프로빗 모형과 로지스틱 회귀모형으로부터는 유사한 적합 결과를 얻게 된다.
로지스틱 회귀모형이 잘 적합되면 프로빗 모형도 잘 적합되고 그 반대도 마찬가지다.
예로, 코골이와 심장병 자료의 경우, 잔차 이탈도는 df=2이고
로짓 연결함수의 경우 그 값이 2.81이고 프로빗 연결함수의 경우 1.87이다.

프로빗 모형의 모수추정값은 로지스틱 회귀모형의 추정값보다 크기가 더 작다.
왜냐하면 이 연결함수 둘 다 확률값을 표준화된 정규분포와 로지스틱 분포의 점수로 변환하지만
이 두 분포가 서로 퍼진 정도가 다르기 때문이다.
즉 표준정규분포는 $\mu = 0$ 이고 $\sigma = 1$ 이다.
표준 로지스틱 분포는 $\mu = 0$ 이고 $\sigma = 1.8$ 이다.
두 모형이 잘 적합될 때에 로지스틱 회귀모형의 모수추정값은 프로빗 모형의 모수추정값의 약 1.8배이다.

5.5.6 이항회귀모형을 위한 CDF와 곡선 형태

잠재변수에 대해 가정한 분포에 따라 이제 설명할 것처럼 P(Y=1)에 대한 곡선의 형태도 결정이 된다.
확률변수 Z에 대한 누적분포함수(CDF) F는 모든 누적확률을 정의하는 함수로 다음과 같다.

$$F(z) = P(Z \leq z), -\infty < z < \infty$$

z의 값이 그 정의된 범위 내에서 증가함에 따라 F(z)는 0에서 1로 증가한다.
Z가 연속 확률변수인 경우, z의 함수로 cdf를 그리게 되면 로지스틱 회귀곡선으로 얻은 것과 같은 S 모양을 갖는다.
F가 어떤 연속확률분포에 대한 cdf라고 하면, 이 관계로부터 P(Y=1)와 설명변수 간에 다음과 같은 관계가 있는
주어지는 이항반응변수 모형을 생각해볼 수 있다.

$$P(Y=1) = F(\alpha + \beta_1 x_1 + \ldots + \beta_p x_p)$$

그러면 F^{-1} 은 선형 예측식을 구하기 위해 P(Y=1)에 적용되는 연결함수이다.
F가 표준정규분포의 cdf일 때 F^{-1} 는 프로빗 연결함수이다.
이 연결함수는 P(Y=1)의 곡선이 선형 예측식의 함수로 정규분포의 cdf 형태를 가지도록 P(Y=1)를 변환하는 역할을 한다.
한 개의 설명변수가 있는 프로빗 모형에서 P(Y=1)DP 대한 정규분포 cdf의 모수(μ, σ)는
프로빗 모형의 모수(α, β)와 $\mu = -\alpha/\beta \quad \sigma = 1/|\beta|$ 관계가 있다.
α 와 $\beta > 0$ 의 선택에 따라 정규분포가 대응된다.

5.5.6 이항회귀모형을 위한 CDF와 곡선 형태

코골이와 심장병 자료의 경우

$$\text{probit}[\hat{P}(Y=1)] = -2.061 + 0.188x$$

이 프로빗 적합 결과는 P(Y=1)가 $\hat{\mu} = -\hat{\alpha}/\hat{\beta} = 11.0$ 이고 $\hat{\sigma} = 1/|\hat{\beta}| = 5.3$인 정규분포 cdf 모양을 따른다는 의미이다.
심장병의 추정확률값 1/2이 되는 코골이 정도는 x = 11.0이다.
이 자료에서 관측된 코골이 수준은 0에서 5까지 제한된 범위값을 가지고 11보다 아주 작기 때문에
이 범위 내에서 구한 확률추정값은 매우 작다.

5.6 로지스틱 회귀모형에 대한 표본크기와 검정력

많은 연구의 목적은 특정한 변수가 이항반응변수에 대해 효과를 미치는지를 알아보는 것이다.
연구계획 단계에서는 이 실질적으로 유의한 크기가 있는 효과를 제대로 탐지할 확률을 높게 만들기 위해 필요한 표본크기를 결정해야 한다.

5.6.1 두 비율을 비교하기 위한 표본크기

두 그룹을 비교하도록 설계된 연구에 대해, 두 그룹의 "성공" 확률 π_1과 π_2가 서로 같다는 가설을 고려해보자.
반응변수에 따라 교차분류된 2X2 분할표에서 고정된 유의수준 α 에 대해
P-값 $\leq \alpha$ 이면 귀무가설을 기각하는 검정을 수행할 수 있다.
표본크기를 결정하기 위해서는 π_1 과 π_2 간에 실제로 일정한 크기의 차이가 존재할 때에,
이 차이를 발견하지 못할 확률 β 을 명시해야 한다.
그러면 α = P(제 1종오류) 이고 β = P(제 2종오류)이다. 검정력은 1- β 이다.

두 그룹에 대해 같은 표본크기를 사용하는 연구에서 근사적으로 다음과 같은 표본크기가 필요하다.

$$n_1 = n_2 = (z_{\alpha/2} + z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]/(\pi_1 - \pi_2)^2$$

이 식을 계산하기 위해서는 π_1과 π_2, α ,β를 알아야한다.
유의 수준 0.05에서 귀무가설을 검정하기 위해 π_1 과 π_2 의 참 값이 0.20과 0.30일때
P(제 2종오류)=0.10을 원한다고하자.
그러면 =0.05, =0.10, z-점수는 $z_{0.025} = 1.96$ 과 $z_{0.10} = 1.28$ 이므로 표본크기는 다음과 같다.

$$\begin{aligned} n_1 = n_2 &= (z_{\alpha/2} + z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]/(\pi_1 - \pi_2)^2 \\ &\approx (1.96 + 1.28)^2[(0.2)(0.8) + (0.3)(0.7)]/(0.2 - 0.3)^2 = 389 \end{aligned}$$

이 식은 또한 대응되는 $\pi_1 - \pi_2$ 의 신뢰구간을 구하는데 필요한 표본크기를 제공한다.
그러면 1- α 는 신뢰구간의 신뢰수준이 되고 β 는 신뢰구간이 효과가 유의하지 않다고 나타낼 확률이다.

5.6.2 로지스틱 회귀모형에서의 표본크기

다음의 로지스틱 회귀모형에서 $H_0 : \beta_1 = 0$ 의 가설검정을 고려해보자.

$$logit[P(Y=1)] = \beta_0 + \beta_1 x$$

x가 질적 변수인 경우 이 모형은 두 성공확률을 비교한다.
x가 양적 변수인 경우엔 적절한 검정력을 확보하기 위해 필요한 표본크기는 x값의 분포에 의존하기 때문에 쉽게 결정하기 어렵다.
x가 확률변수이고 정규분포를 따른다고 가정하자.
x의 평균값에서의 성공 확률을 π 라고 하자. 필요한 표본크기는 x가 평균값보다 표준편차만큼 증가했을 때의 성공확률과 π 를 비교하는 오즈비 θ 에 의해 결정된다.
$\lambda = log(\theta)$ 라고 나타내면 단측검정에 필요한 표본크기의 근사값은 다음과 같다.

$$n = [z_\alpha + z_\beta exp(-\lambda^2/4)]^2(1+2\pi\delta)/(\pi\lambda^2)$$

여기서

$$\delta = [1+(1+\lambda^2)exp(5\lambda^2/4)]/[1+exp(-\lambda^2/4)]$$

이다.
π 가 0.05에 가까워지고 | λ |가 귀무가설값 0에서 멀어짐에 따라 표본크기 n은 점점 감소한다.

5.6.2 로지스틱 회귀모형에서의 표본크기

다중 로지스틱 회귀모형에서 같은 크기의 부분 효과(partial effect)를 검정하기 위해서는 표본크기 n이 더 커야 한다.
R이 관심 있는 예측변수 x_j 와 모형에 있는 다른 예측변수 간의 다중상관관계를 나타낸다고 하자.
여기에서의 표본 크기는 다음과 같다.

$$n = [z_\alpha + z_\beta exp(-\lambda^2/4)]^2(1+2\pi\delta)/(\pi\lambda^2) \Rightarrow n/(1-R^2)$$

이 식에서 π 는 모든 설명변수들의 평균값에서의 확률값이고 오즈비 θ 는 다른 예측변수들이 평균값을 가질 때에 관심 있는 예측변수 x_j 의 효과를 나타낸다.
그러나 결과는 사용이 제한적이다.
왜냐면 다중 로지스틱 회귀모형에서는 R=0인 경우조차 한 효과는 변수들이 추가되면
그 효과 크기가 변하기 때문이다.

지금까지 살펴본 식들은 표본크기에 대한 대략적인 지표일 뿐이다.
대부분의 실제 문제에서는 π 와 θ 와 R의 대략적인 값만 알 뿐, 설명변수도 정규분포를 따르지 않을 수 있다.

5.6.3 예제. 심장병 발병확률 모형화

어느 중년 모집단을 대상으로 x = 콜레스테롤 수준이 심장병의 발병 확률에 미치는 영향을 나타내기 위한 모형을 세우기 위한 연구를 계획해보자.
과거의 연구 결과 평균 콜레스테롤 수준에서 심장병이 발병할 확률이 0.08로 알려졌다고 가정하자.
연구자들은 다음과 같은 가설에 관심이 있다고 하자.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

콜레스테롤 수준이 증가함에 따라 위험이 증가한다는 대립가설에 대한 검정을 고려해보자.
특히 콜레스테롤이 표준편차만큼 증가했을 때 심장병의 발병 확률이 50% 증가하는지를 검정하는 데 관심이 있다고 하자.
평균 콜레스테롤 수준에서의 심장병이 발병할 오즈는 0.08/0.92 = 0.087이고
평균보다 표준편차만큼 큰 콜레스테롤 수준에서의 오즈는 0.12/0.88 = 0.136이다.
오즈비는 $\theta = 0.136/0.087 = 1.57$ 이므로 $\lambda = log(1.57) = 0.450$ 이고 $\delta = 1.306$ 이다.

$\alpha = 0.05$ 이고 $\beta = 0.10$일 때 $z_\alpha = z_{0.05} = 1.645$ $z_\beta = z_{0.10} = 1.28$ 이다.
따라서 이 연구에서 필요한 표본크기 n = 612이다.

끝 ~~!