

(그림으로 이해하는) 닥터 배의 술술 보건의학통계

Part_2 중급 보건의학통계 맛보기
라 가 영



09. 질병의 위험인자에 대한 연구

- 가. 로지스틱 회귀분석
 - 로짓 변환
 - 로지스틱 회귀모형의 추정
 - 질병이 있을 확률
 - 분류표의 작성
 - 추정된 모형의 유의성 검정
 - 로지스틱 회귀분석에서의 교차비
 - 가변수의 설정
 - Hosmer-Lemeshow's goodness-of-fit test

09. 질병의 위험인자에 대한 연구

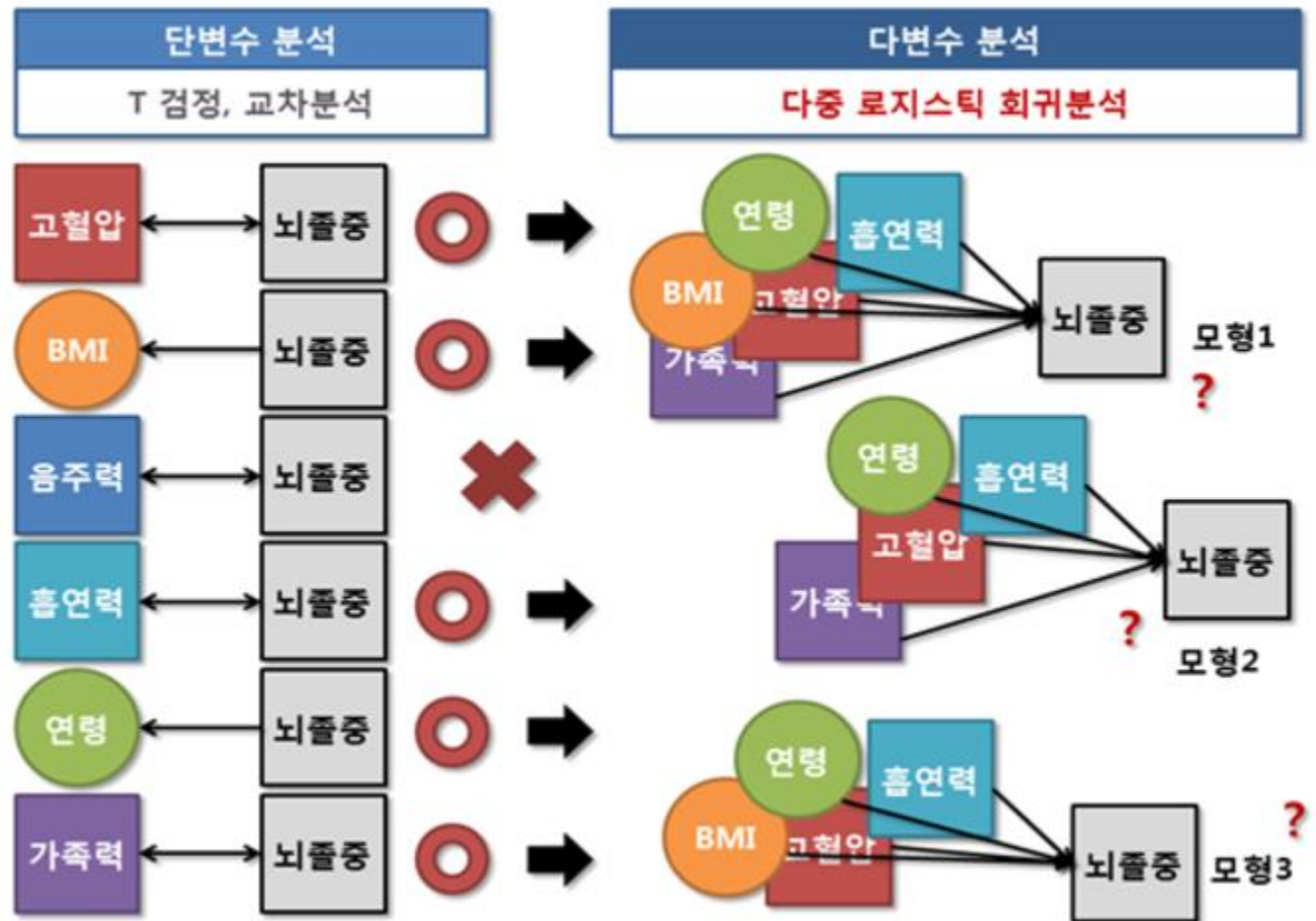
- 선형회귀분석(linear regression analysis) : 혈압이나 혈당을 예측하고, 이에 영향을 미치는 요인을 밝히는 통계적 방법
- 로지스틱 회귀분석(logistic regression analysis) : 특정 질병의 유무에 영향을 미치는 요인을 밝히는 통계적 방법
- 교란변수의 영향을 제거하고 여러 위험인자들이 관련되는 정도를 하나의 모형으로 설명하기 위해서는 로지스틱 회귀분석의 도움이 필요

09. 질병의 위험인자에 대한 연구

영향을 주는 변수 영향을 받는 변수 통계분석방법

범주형 자료	범주형 자료	카이제곱 검정
	연속형 자료	T검정 분산분석
연속형 자료	연속형 자료	회귀분석 구조방정식
	범주형 자료	로지스틱 회귀분석

09. 질병의 위험인자에 대한 연구



교차분석, T 검정 -> 질병의 유무와 개별 독립변수의 관계를 설명
로지스틱 회귀분석 -> 여러 위험인자들이 관련되는 정도를 하나의 모형으로 설명

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 질환의 유무는 범주형 자료이기 때문에 회귀분석에 그대로 적용할 수가 없음
- 일반화 선형모형(*generalized linear model*) : 회귀분석을 확장하여 종속변수 y 를 $f(x)$ 라는 함수로 치환. 보다 폭넓은 현상들을 회귀모형으로 설명하는 것이 가능

일반선형모형

$$y = \alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

일반화 선형모형

$$f(x) = \alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

로지스틱 회귀분석

$$\ln \frac{p}{1-p} = \alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

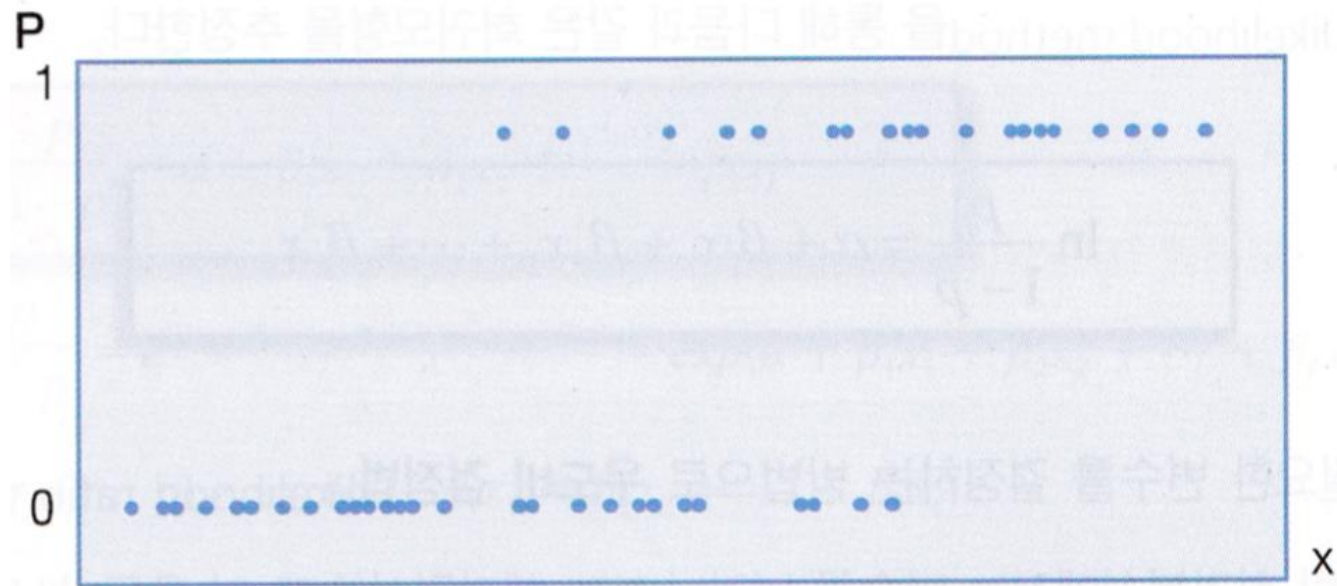
p 는 질환이 있을 확률

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로짓 변환

- E_x 체중(x)과 고혈압 유무(p)의 관계 \rightarrow 질병의 유무는 없거나($p = 0$) 있는($p = 1$) 2가지 값만 취함

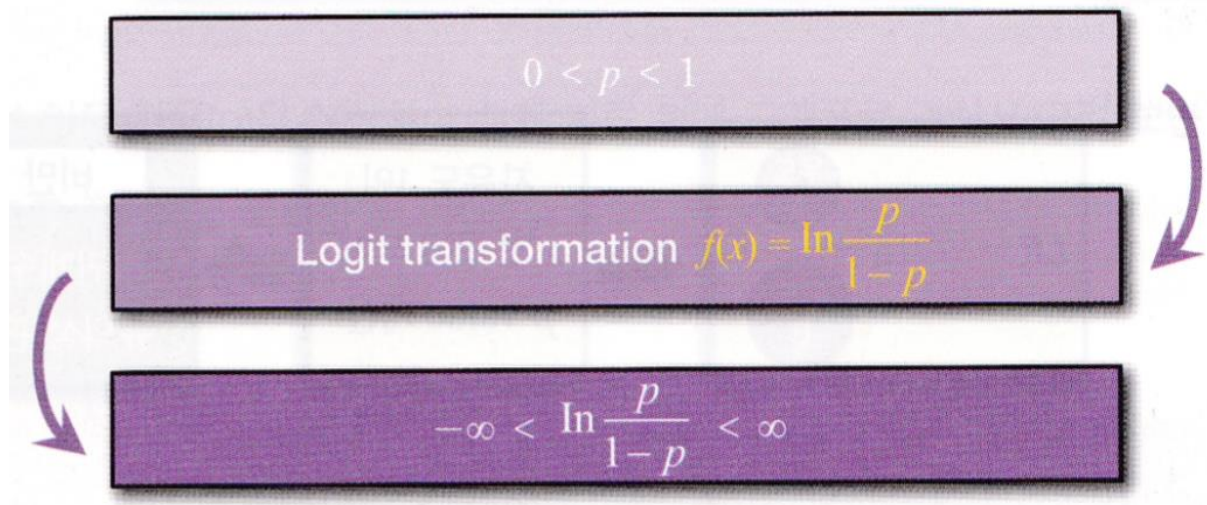


09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로짓 변환

- 위험요인(체중)의 각 수준에 따른 질병(고혈압)이 있을 확률(p)
- 로짓 변환($f(x) = \ln \frac{p}{1-p}$, logit transformation)해주면 $\text{logit } p (= \ln \frac{p}{1-p})$ 값은 $-\infty \sim \infty$ 의 연속형 변수의 형태를 띠게 됨

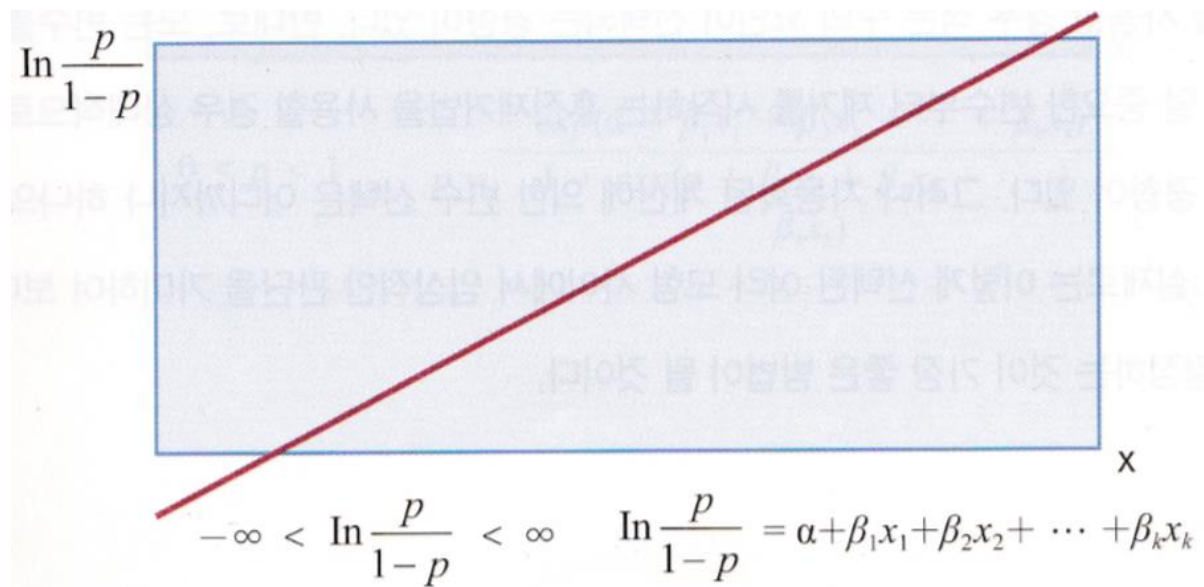


09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로짓 변환

- 로지스틱 회귀분석(logistic regression analysis) : logit p 를 대상으로 회귀분석을 적용한 것



09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로지스틱 회귀모형의 추정

- 최소제곱법 (least square method) : 선형회귀분석에서의 회귀모형 추정법
- 최대우도법 (maximum likelihood method) : 이분형 변수를 대상으로 하는 로지스틱 회귀분석법에서의 회귀모형 추정법

$$\ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로지스틱 회귀모형의 추정

- 변수를 넣거나(전진) 빼면서(후진) 우도비(LR)를 통해 조건에 맞는 변수들을 선택



우도비 검정(likelihood ratio test) : 두 개의 우도의 비를 계산해서 두 모형의 우도가 유의한 차이가 있는지 비교하는 방법

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 질병이 있을 확률

$$\ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$\Rightarrow \frac{p}{1-p} = e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)} = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

$$\Rightarrow p = (1-p) \times \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

$$\Rightarrow p = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - p \times \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

$$\Rightarrow p + p \times \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

$$\Rightarrow p(1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

$$\Rightarrow p = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

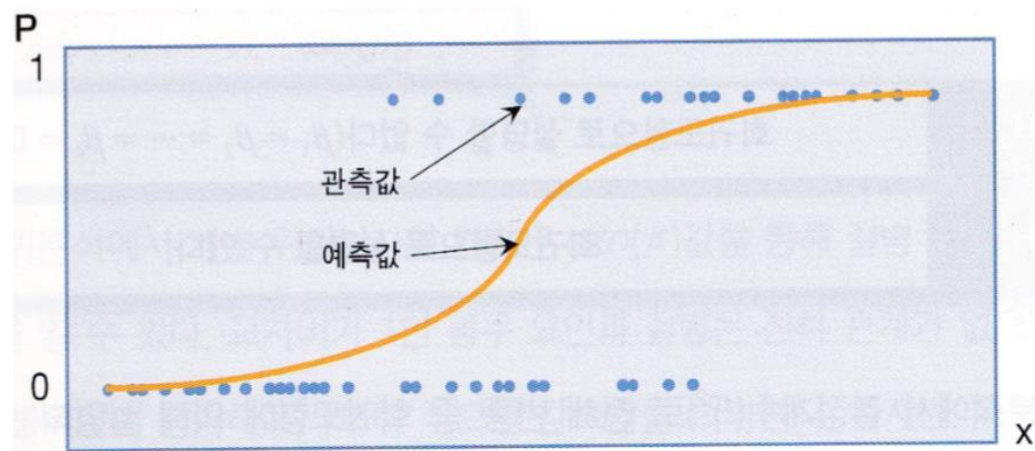
최대우도법을 통해 로지스틱 회귀식이 추정되면, 질병이 있을 확률(p)은 다음과 같이 유도할 수 있다.

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로짓 변환

- 질병이 있을 확률(p)을 실제 자료 위에 겹쳐서 그려보면 S자 모양으로 로지스틱 곡선으로 표현됨



$$0 < p < 1 \quad p = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 분류표의 작성

- 질병이 있을 확률이 0.5이상인 경우 질병이 있을 것으로 예측하고 0.5 미만인 경우 질병이 없을 것으로 예측
- 몇 %에서 예측과 일치하는지를 조사
- 추정된 회귀모형을 평가하는 한 방법

	질병 미발생 예측 ($p_i \geq 0.5$)	질병 발생 예측 ($p_i < 0.5$)	분류 정확도
질병 미발생	117	13	90.0%
질병 발생	38	21	35.6%
			73.0%

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 추정된 모형의 유의성 검정

선형회귀분석	로지스틱 회귀분석
F-검정을 이용하여 유의성 검정	모형계수 전체 테스트(model chi-square test)를 통해 유의성 검정
결정계수(R^2)로 전체 변동 중 회귀모형에 의해 설명되는 변동의 비율을 설명	Cox & Snell의 결정계수(R^2_{CS})와 Nagelkerke의 결정계수(R^2_N)를 통해 회귀식의 설명력을 제시

귀무가설 H_0	회귀모형으로 설명할 수 없다($\beta_1 = \beta_2 = \dots = \beta_k = 0$).
대립가설 H_1	회귀모형으로 설명할 수 있다.

R^2_N 이 더 널리 쓰이는데, 선형회귀분석의 결정계수보다 다소 작은 값을 갖는 경향이 있음

09. 질병의 위험인자에 대한 연구




가. 로지스틱 회귀분석

- 로지스틱 회귀분석에서의 교차비

환자대조군 연구 (후향적)

위험인자 유/무

질병발생 유/무

	질병 발생	질병 미발생	
위험인자 있음	A	B	
위험인자 없음	C	D	

질병이 있을 때 위험인자의 $Odds = \frac{A}{C}$

질병이 없을 때 위험인자의 $Odds = \frac{B}{D}$

교차비(Odds ratio) : 질병이 있는 경우 위험인자 유무의 비와 질병이 없는 경우 위험인자 유무의 비의 비

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로지스틱 회귀분석에서의 교차비

0과 1로 코딩	1과 2로 코딩		질병 발생	질병 미발생
$x_1 = 1$	$x_1 = 2$	위험인자 있음	p_1	$1 - p_1$
$x_1 = 0$	$x_1 = 1$	위험인자 없음	p	$1 - p$

$$\ln \frac{p_1}{1 - p_1} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

$$\ln \frac{p}{1 - p} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

위험인자1을 제외하고
나머지 위험인자는 모두 같으므로
 $x_2 = x_2, \dots, x_k = x_k$

➔

 $\ln \frac{p_1}{1 - p_1} - \ln \frac{p}{1 - p} = \beta_1 (x_1' - x_1) = \beta_1 \times 1 = \beta_1$

➔

$\ln \frac{\frac{p_1}{1 - p_1}}{\frac{p}{1 - p}} = \beta_1$

$odds\ ratio = \frac{\frac{p_1}{1 - p_1}}{\frac{p}{1 - p}}$

➔

 $\ln(odds\ ratio) = \beta_1$

➔

$odds\ ratio = e^{\beta_1} = \exp(\beta_1)$

다른 위험인자(x_2, \dots, x_k)는 모두 동일하다고 가정할 때 위험인자 x_1 이 질병유무와 관련된 정도가 회귀계수(β_1)를 통해 어떻게 표현되는지

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로지스틱 회귀분석에서의 교차비

- 개별 위험인자의 영향은 $\exp(\beta)$ 값을 통해 질병 발생에 대한 교차비(odds ratio)로 표현됨
- 교차비가 1인 경우 요인과 질병은 전혀 관계가 없음
- 1보다 크면 요인에 의해 질병의 위험이 증가하고, 1보다 작으면 감소함을 의미
- 교차비의 95% 신뢰구간이 더 많은 정보를 제공해 줄 수 있기 때문에 교차비, 유의수준, 95% 신뢰구간을 동시에 제시 하는 것이 좋음

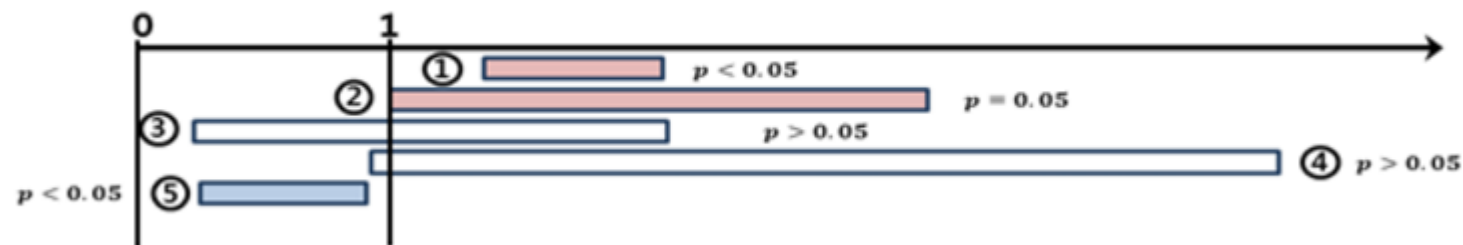
09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 로지스틱 회귀분석에서의 교차비

- 1의 경우 95% 신뢰구간이 1보다 큰 구간에 위치. 실제 교차비가 1이어서 질병과 위험인자 사이에 연관성이 전혀 없을 가능성은 5% 미만($p < 0.05$)
- 5의 경우 95% 신뢰구간이 0과 1사이에 완전하기 위치. 위험인자가 있는 경우 질병의 위험도가 통계적으로 유의하게 낮음($p < 0.05$)

교차비(Odds ratio)의 95% 신뢰구간



09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 가변수의 설정

- 범주가 3개 이상으로 나뉘는 요인(1, 2, 3, ... 으로 코딩 된 경우)의 경우에는 기준으로 삼을 참조변수(reference)를 정하고 가변수(dummy variables)를 통해 위험도를 분석
- 범주의 수가 n 개라면, 가장 낮은 값으로 코딩 된 범주를 참조치로 정하고 가변수의 개수(비교의 횟수)는 $n-1$ 개

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- 가변수의 설정

자료 코딩		$\ln \frac{p_i}{1-p_i} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots$		
Reference	20대	0	0	0
	30대	1	0	0
	40대	0	1	0
	50대 이상	0	0	1

결과 해석		Reference	
30대	의	20대	에 대한 질병 유무의 교차비는 $\exp(\beta_1)$ 이고,
40대	의	20대	에 대한 질병 유무의 교차비는 $\exp(\beta_2)$ 이고,
50대 이상	의	20대	에 대한 질병 유무의 교차비는 $\exp(\beta_3)$ 이다.

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- Hosmer-Lemeshow's goodness-of-fit test

- Hosmer-Lemeshow's goodness-of-fit test : 로지스틱 회귀분석에서 모형의 적합도를 평가하는 방법 중 하나
- 이 검정에서는 귀무가설이 모형이 적합하다 이므로 $p \text{ value} > 0.05$ 일 때 모형이 적합하다고 해석
- 표본수가 충분히 클 때만 적용 가능

귀무가설 H_0	모형은 적합하다.
대립가설 H_1	모형이 적합하지 않다.

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- Hosmer-Lemeshow's
goodness-of-fit test

- 로지스틱 회귀분석으로 추정된 모형을 평가하는 여러 방법들
 - 모형계수 전체 테스트 : 회귀모형의 유의성 검정은 물론, 분류표를 통해 회귀모형이 현상을 얼마나 잘 예측하고 있는지 파악 가능
 - Nagelkerke의 결정계수(R^2_N) : 회귀식의 설명력을 볼 수 있음
 - Hosmer-Lemeshow의 모형적합도 검정법

09. 질병의 위험인자에 대한 연구

가. 로지스틱 회귀분석

- Hosmer-Lemeshow's goodness-of-fit test

	선형 회귀분석	로지스틱 회귀분석
종속변수	연속형	범주형
회귀모형	$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$	$\ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$
회귀모형의 추정방법	최소제곱법 (Least square method)	최대우도법 (Maximum likelihood method)
회귀모형의 유의성 ($H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$)	최소제곱법 (F 검정)	모형 계수 전체 테스트 (Model chi-square test)
회귀계수의 유의성 ($H_0: \beta_1 = 0$)	T 분포를 이용	우도비 검정 (Likelihood ratio test)
회귀모형의 평가	R^2	1. 분류표 2. Nagelkerke의 R^2 3. Hosmer-Lemeshow 검정
얼고자 하는 값	회귀계수(β)	교차비($\exp(\beta)$)의 95% 신뢰구간