로지스틱 회귀분석

정의

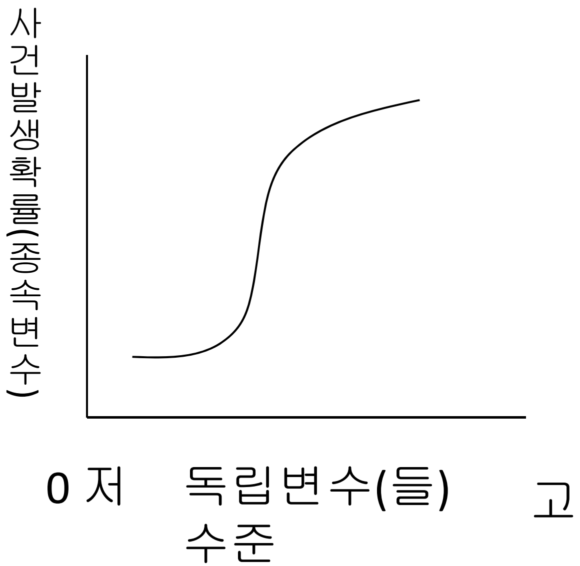
- 로지스틱 회귀분석 (logitic regression)은 어떤 사건이 발생하는지 안하는지를 직접 예측하는 것이 아니라,그 사건이 발생할 확률을 예측한다. 일반적으로 종속변수의 범주가 두 개인 경우에 적용된다.
- 독립변수와 종속변수의 관계를 단순회귀분석과 다중회귀분석은 선형으로 가정하는데 비해, 로지스틱 회귀분석은 S자형으로 가정한다.

로지스틱 회귀분석의 예

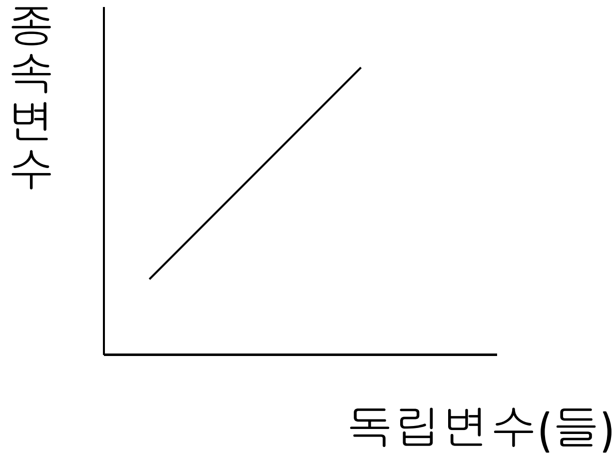
- 기업의 크기에 따라 노조를 가지고 있는지 여부 조사
- 기혼 여성의 경제활동 참가여부를 나이, 자녀의 수, 남편의 연봉으로 모형화
- 책임보험 가입 여부를 가장의 나이, 유동자산규모, 가장의 직업으로 설명
- 심장질환 발생 여부를 환자의 나이, 성별, 흡연여부, 콜레스테롤 수치, 혈압의 함수로 설명

독립변수와 종속변수 관계에 대한 가정

- 로지스틱 회귀분석



- 단순회귀분석과 다중회귀분석



자료

- 종속변수 :명목척도
- 독립변수:명목척도,간격척도, 비율척도
- 독립변수가 명목척도로 측정된 경우 일반적 회귀 분석에서 처럼 더미변수로 변경하여 입력한다.

Response Function

- 종속변수가 0 or 1을 갖는 binary

Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

- 단순회귀분석을 고려하면

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad E(\epsilon_i) = 0$$

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

종속변수가 binary 인 경우의 문제점

- Nonnormal Error Term
  - ϵ_i이 두 가지 값만 가능
  - $Y_i = 1: \epsilon_i = 1 - \beta_0 - \beta_1 X_i$
  - $Y_i = 0: \epsilon_i = -\beta_0 - \beta_1 X_i$
- Nonconstant Error Variance

$$\sigma^2(\epsilon_i) = \pi_i(1 - \pi_i)$$
$$= (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$$

Logistic Regression Model

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

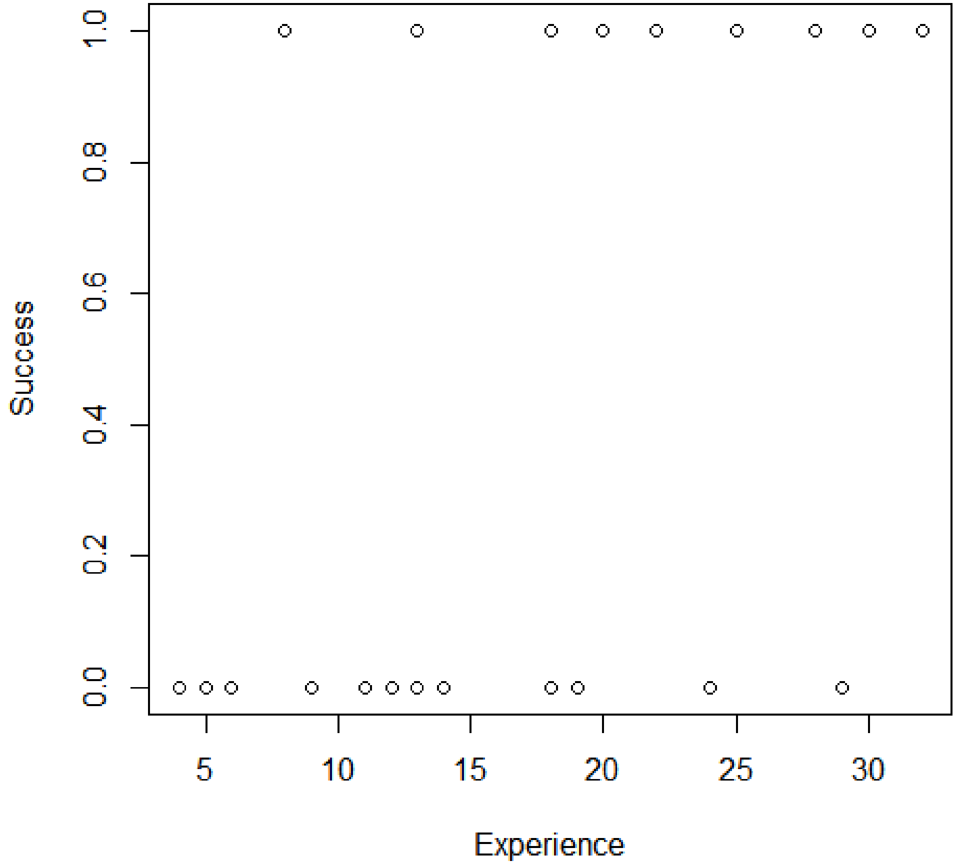
$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_i$$

Odds: 성공확률/실패확률

Example: Programming Experience

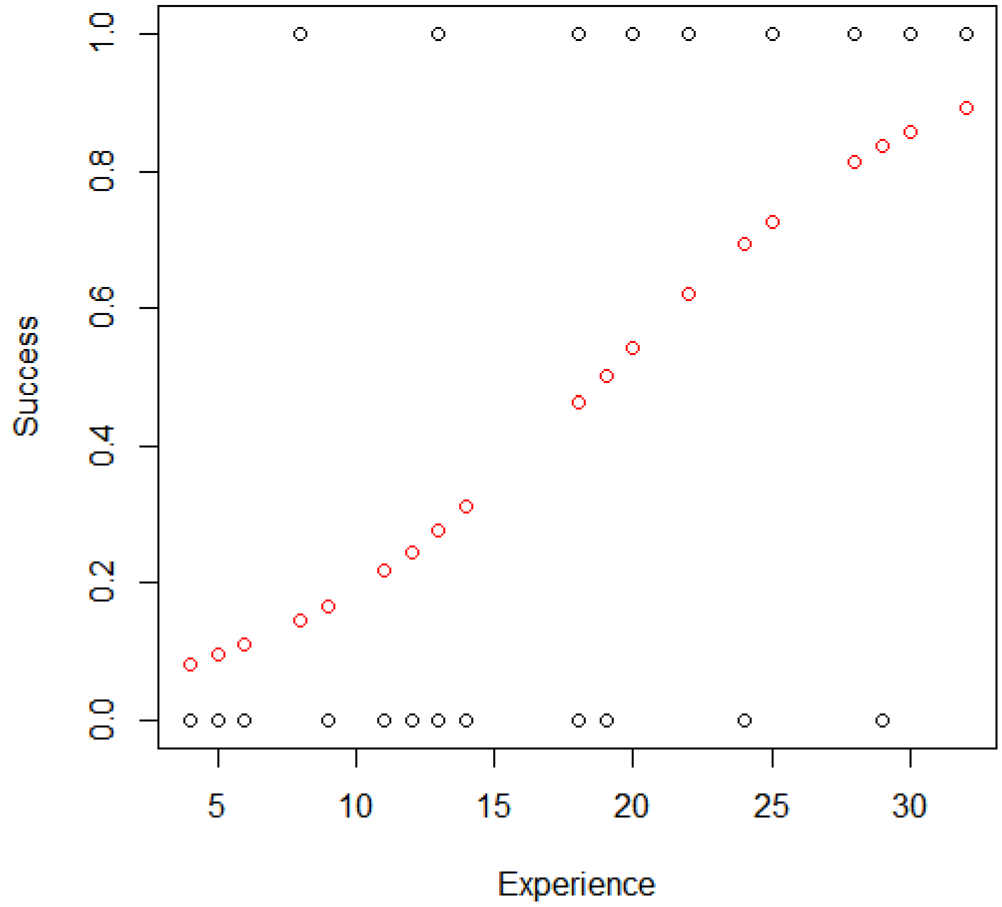
- 컴퓨터 프로그래밍 경험이 특정한 분석 문제를 완성하는데 영향이 있는지 분석

	Experience	Success
1	14	0
2	29	0
3	6	0
4	25	1
5	18	1
6	4	0
7	18	0
8	12	0
9	22	1
10	6	0



$$\hat{\pi} = \frac{\exp(-3.06 + 0.16X)}{1 + \exp(-3.06 + 0.16X)}$$

```
> data=read.csv("programming.csv")
> model=glm(Success~Experience,data=data,family=binomial(logit))
> summary(model)

Call:
glm(formula = Success ~ Experience, family = binomial(logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8992  -0.7509  -0.4140   0.7992   1.9624

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
Experience   0.16149    0.06498   2.485   0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.296  on 24  degrees of freedom
Residual deviance: 25.425  on 23  degrees of freedom
AIC: 29.425

Number of Fisher Scoring iterations: 4
```



$X_1 = 14$에 대한 예측치: $\widehat{\pi_1} = \frac{\exp(-3.06+0.16(14))}{1+\exp(-3.06+0.16(14))} = 0.310$

➔ 경력이 14년인 프로그래머가 특정 작업을 마칠 확률은 0.31이다.

b_1의 해석

$$\log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = b_0 + b_1 X_i$$

- b_1: X가 1 증가할 때 $\log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i})$의 증가분

$$b_1 = \log(Odds_1) - \log(Odds_2)$$

$$= \log\left(\frac{Odds_1}{Odds_2}\right)$$

$$\textbf{Odds Ratio}(\widehat{OR}) = \frac{Odds_1}{Odds_2} = \exp(b_1)$$

b_1의 해석

$$\widehat{OR} = \exp(0.1615) = 1.175$$

- **<u>경험이 1개월 증가할 때 특정 작업을 완료할 Odds가 17.5% 증가한다.</u>**

- 10개월 경력자와 25개월 경력자의 차이?

$$\widehat{OR} = \exp(15(0.1615)) = 11.3$$

➔ 특정작업을 완료할 Odds가 11배 증가한다.

반복 측정된 자료

- 한 개의 X값에서 여러 개의 Y가 측정된 경우
- X_i에서의 관측치가 0, 1이 아니라 n_i개 중 $Y_{.i}$개의 성공 관측
- Binomial Distribution

$$f(Y_{.i}) = \frac{n_i!}{Y_{.i}!\,(n_i - Y_{.i})!} \pi_i^{Y_{.i}} (1 - \pi_i)^{n_i - Y_{.i}}$$

Example: Coupon Effectiveness

- 가격을 할인해 주는 쿠폰의 효과를 검증하기 위해 무작위로 추출된 각 200개의 가구에 5,10,15,20,30 달러의 쿠폰을 제공했다.

	Price_reduc	N	N_redeemed
1	5	200	30
2	10	200	55
3	15	200	70
4	20	200	100
5	30	200	137

```
> data=read.csv("coupon.csv")
> model2=glm(cbind(N_redeemed,N-N_redeemed)~Price_reduc,data=data,family=binomial(logit))
> summary(model2)

Call:
glm(formula = cbind(N_redeemed, N - N_redeemed) ~ Price_reduc,
    family = binomial(logit), data = data)

Deviance Residuals:
      1        2        3        4        5
-0.8988   0.6677  -0.1837   0.7612  -0.5477

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.044348   0.160977  -12.70   <2e-16 ***
Price_reduc  0.096834   0.008549   11.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149.4627  on 4  degrees of freedom
Residual deviance:   2.1668  on 3  degrees of freedom
AIC: 33.793

Number of Fisher Scoring iterations: 3
```

$$\hat{\pi} = \frac{\exp(-2.04 + 0.0968X)}{1 + \exp(-2.04 + 0.0968X)}$$

$\widehat{OR} = \exp(0.0968) = 1.102$ ➔ 쿠폰의 할인액이 1달러 증가할 때 쿠폰을 사용할 Odds가 10% 증가한다.

Multiple Logistic Regression

$$E(Y\quad) = \pi_i$$
$$= \frac{\exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}$$

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Example: Disease Outbreak

- 모기에 의한 유행병의 전염을 연구하기 위해 두 지역에서 최근에 병에 걸린 사람들을 무작위 추출했다. 특정 증상을 보였는지 여부를 아래의 설명변수로 모형화한다.
  - 나이 (X1)
  - 사회경제적 위치 (X2=1 if Middle, X3=1 if Lower)
  - 지역 (X4=0 for sector 1, X4=1 for sector 2)

	case	age	status_middle	status_lower	sector	disease
1	1	33	0	0	0	0
2	2	35	0	0	0	0
3	3	6	0	0	0	0
4	4	60	0	0	0	0
5	5	18	0	1	0	1
6	6	26	0	1	0	0
7	7	6	0	1	0	0
8	8	31	1	0	0	1
9	9	26	1	0	0	1
10	10	37	1	0	0	0

```
> data=read.csv("disease.csv")
> model3=glm(disease~.,data=data,family=binomial(logit))
> summary(model3)

Call:
glm(formula = disease ~ ., family = binomial(logit), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6552  -0.7529  -0.4788   0.8558   2.0977

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.31293    0.64259  -3.599 0.000319 ***
age            0.02975    0.01350   2.203 0.027577 *
status_middle  0.40879    0.59900   0.682 0.494954
status_lower  -0.30525    0.60413  -0.505 0.613362
sector         1.57475    0.50162   3.139 0.001693 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.32  on 97  degrees of freedom
Residual deviance: 101.05  on 93  degrees of freedom
AIC: 111.05

Number of Fisher Scoring iterations: 4
```

$$\hat{\pi} = \frac{\exp(-2.31 + 0.0297X_1 + 0.409X_2 - 0.305X_3 + 1.57X_4)}{1 + \exp(-2.31 + 0.0297X_1 + 0.409X_2 - 0.305X_3 + 1.57X_4)}$$

계수추정치의 해석

	Estimated Coefficients	Estimated Odds Ratio
b_1	0.02975	1.030
b_2	0.4088	1.505
b_3	-0.3053	0.737
b_4	1.5747	4.829

- 사회경제적 위치와 지역이 주어져 있을 때 나이가 1살 많아지면 특정 증상을 가질 Odds는 3% 증가한다.
- 사회경제적 위치와 나이가 주어져 있을 때 sector 2 지역 주민의 Odds는 약 5배 sector 1 주민에 비해 약 5배 크다.

각 계수에 대한 유의성 검정: Wald Test

$$H_0: \beta_k = 0$$
$$H_a: \beta_k \neq 0$$

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.31293    0.64259  -3.599 0.000319 ***
age             0.02975    0.01350   2.203 0.027577 *
status_middle   0.40879    0.59900   0.682 0.494954
status_lower   -0.30525    0.60413  -0.505 0.613362
sector          1.57475    0.50162   3.139 0.001693 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 각 설명변수의 영향이 유의한지 한번에 검정
  - 다중회귀분석의 t-test에 해당

모형비교: Deviance Goodness-of-fit Test

```
> model3=glm(disease~.,data=data,family=binomial(logit))
> model4=glm(disease~age+sector,data=data,family=binomial(logit))
> anova(model3,model4,test="Chisq")
Analysis of Deviance Table

Model 1: disease ~ age + status_middle + status_lower + sector
Model 2: disease ~ age + sector
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        93     101.05
2        95     102.26 -2  -1.2052   0.5474
```

- Reduced Model과 Full Model의 차이가 유의한지 검정
- 여러 설명변수가 주는 영향이 유의한지 한번에 검정
  - 다중회귀분석의 F-test와 유사

새 관측치에 대한 예측

- $\widehat{\pi_h}$가 크면 1로 예측
- $\widehat{\pi_h}$가 작으면 0으로 예측
- Cutoff point의 결정
  - 0.5
  - 여러 point를 시도한 후 best 선택
  - 사전 지식에 의한 선택

Example: Disease Outbreak

```
> cbind(data$disease,model4$fitted)
   [,1]       [,2]
1     0 0.20284900
2     0 0.21248645
3     0 0.10345506
4     0 0.35944968
5     1 0.14088829
6     0 0.17170208
7     0 0.10345506
8     1 0.19354123
9     1 0.17170208
10    0 0.22245400
11    0 0.15956318
12    0 0.15956318
13    0 0.17590786
```

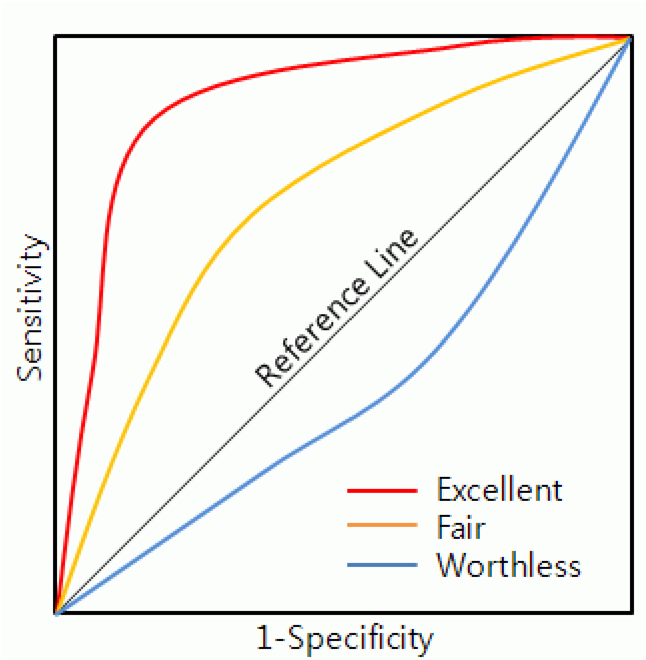
- 98명 중 31명이 증상
- Cutoff=31/98=0.316

```
> xtabs(~data$disease+(model4$fitted>0.316))
            model4$fitted > 0.316
data$disease FALSE TRUE
           0    47   20
           1     8   23
```

- **민감도 (Sensitivity)**: True를 True로 구분한 비율 = 23/31=0.74
- **특이도 (Specificity)**: False를 False로 구분한 비율 = 47/67=0.70
- Error rate=28/98=0.29

ROC (Receiver Operating Characteristic) Curve

- y축: 민감도
- x축: 1- 특이도=False positive



- ROC curve의 아래쪽 면적(AUC)이 클수록 좋은 모형
- 왼쪽 코너에 가까운 포인트를 Cutoff로 정하는 것도 한 방법

Example: ROC curve for Disease Outbreak