

# 5장 데이터 분석기법(회귀분석)

- 회귀분석
  - 선형회귀분석
    - 단순선형회귀분석
    - 다중선형회귀분석
  - 로지스틱 회귀분석
- 교차 분석

# 01 회귀분석의 종류

- **선형 회귀분석(Linear Regression)**: 일반적으로 사용하는 모델링 기술 중 하나이며 종속 변수( $y$ )는 연속적이며, 독립 변수( $x$ )는 연속적이거나 이산적일 수 있다. 회귀선은 선형을 가진다.
  - 단순선형 회귀(simple linear regression): 독립변수가 1개 존재
  - 다중선형회귀(multiple linear regression): 독립변수가 2개 이상 존재

Simple  
Linear  
Regression

$$y = b_0 + b_1 * x_1$$

Multiple  
Linear  
Regression

Dependent variable (DV)      Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

Constant      Coefficients

- **로지스틱 회귀분석(Logistic Regression)**: 종속변수( $y$ )가 이진(1/0, 참/거짓, 예/아니요)일 때 주로 사용되는 회귀 분석이다.
- **비선형 회귀분석(Nonlinear Regression)**

# 02-1 선형회귀분석: 단순선형회귀 분석의 예

- 하루최고기온으로부터 음료지불금액을 예측

월	1	2	3	4	5	6
하루최고기온(℃)	9.1	10.2	14.1	19.8	25.0	26.8
음료지출금액(엔)	3416	3549	4639	3857	3989	4837
월	7	8	9	10	11	12
하루최고기온(℃)	31.1	34.0	28.5	22.9	15.7	11.3
음료지출금액(엔)	5419	5548	4311	4692	3607	4002

$$y = b_0 + b_1 \times x_1$$

- 최소제곱법을 사용해 회귀직선을 구함

$$\hat{y} = 2947.8 + 66.4 \times x$$

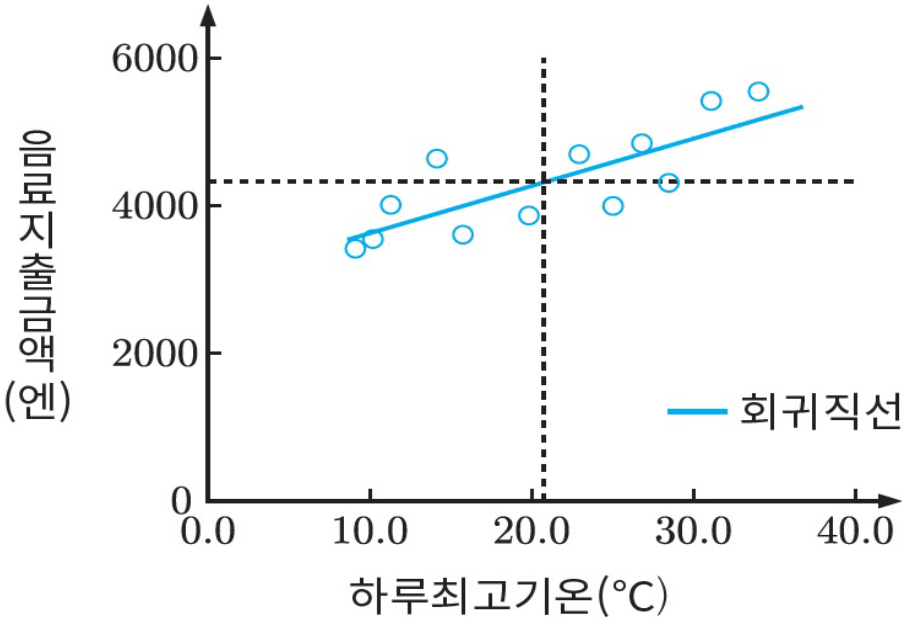
하루최고기온: 9.1~34.0℃

하루최고기온이 10℃인 경우 음료지불금액

:  $2947.8 + 66.4 \times 10 = 3611.8$

기온이 1℃ 상승하면 평균적으로 음료지불금액이 66.4엔씩 증가

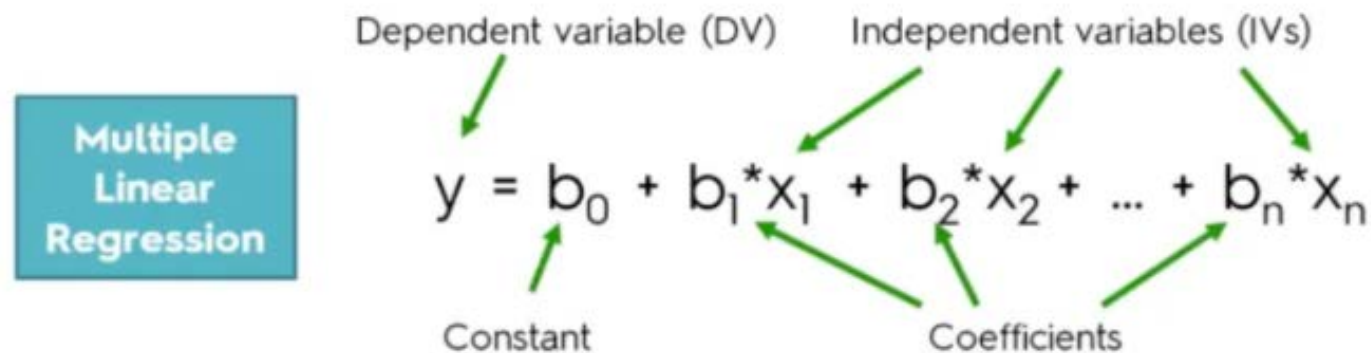
회귀직선은 각 변수의 평균값을 좌표로 갖는 점(20.7, 4322.2)를 통과



$$b_0 = \bar{y} - b_1 \bar{x}$$

## 02-2 선형회귀분석: 다중 선형회귀 분석

- 종속 변수의 변화가 하나의 독립변수만으로 충분히 설명할 수 없는 경우가 많음. 따라서 독립변수를 적절히 여러 개 선택하여 이들의 함수로서 종속변수를 설명하는 것이 더 정확할 수 있음.
- 이 경우의 모델을 다중선형회귀 모델이라 한다.



The diagram illustrates the Multiple Linear Regression equation:  $y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$ . A blue box on the left contains the text "Multiple Linear Regression". Labels with green arrows point to various parts of the equation: "Dependent variable (DV)" points to  $y$ ; "Independent variables (IVs)" points to  $x_1, x_2, \dots, x_n$ ; "Constant" points to  $b_0$ ; and "Coefficients" points to  $b_1, b_2, \dots, b_n$ .

- 여기에서,  $b_0, b_1, b_2, \dots, b_n$ 은 구해야 할 회귀계수이다.
- 최소제곱법(least square method)에 의해 회귀계수를 구할 수 있다.
- 각 회귀계수에 대한 검정도 단순회귀분석의 경우와 동일하게 수행한다. (예: 결정계수)

## 02-2 선형회귀분석: 다중 선형회귀 분석 (예제)

- "더운 날에는 아이스크림이 많이 팔릴 것이다"

→ 아이스크림 판매수량과 최고기온 데이터를 조사하여 회귀분석을 수행

$$\hat{y} = 210.8 + 134.2x$$

→ 예상최고기온이 30 °C 라면 예상 판매수량은 4236.8개

최고기온이 1°C 상승하면,  
아이스크림의 판매수량은  
134.2개씩 증가

- "가격을 저렴하게 하면 많이 판매된다"

→ 회귀식의 오른쪽 항에 변수를 추가

- "평일보다는 휴일에 많이 팔린다"

→ 더미변수(dummy variable): 휴일인 경우에는 1, 평일인 경우에는 0

- 회귀직선:  $\hat{y} = 195.4 + 118.1x - 5.8p + 30.4D$

p: 아이스크림 1개의 가격(엔)

D: 더미변수(휴일이면 1, 평일이면 0)

"아이스크림의 가격을 1엔 올리면 매상이 5.8개 줄어든다"

"휴일은 평일에 비해서 매상이 30.4개 올라간다"

## 02-2 선형회귀분석: 다중 선형회귀 분석

The diagram shows the equation  $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$ . Green arrows point from labels to parts of the equation: 'Dependent variable (DV)' points to  $y$ , 'Independent variables (IVs)' points to  $x_1, x_2, \dots, x_n$ , 'Constant' points to  $b_0$ , and 'Coefficients' points to  $b_1, b_2, \dots, b_n$ .

- 회귀모델에 포함시키는 독립변수의 선정 기준

- 종속변수와 높은 상관관계를 갖는다.
- 선택된 독립변수들은 서로 상호간에 낮은 상관관계를 갖는다. (다중공선성 문제 회피)
- 독립변수의 개수는 적을수록 좋다

❖ **다중공선성(multicollinearity):** 독립변수들 간에 밀접한 상관관계가 존재하는 것을 말하며, 이와 같은 경우에는 독립변수의 계수가 정확히 추정되지 못하는 문제가 발생함.

- 독립변수의 선택 방법

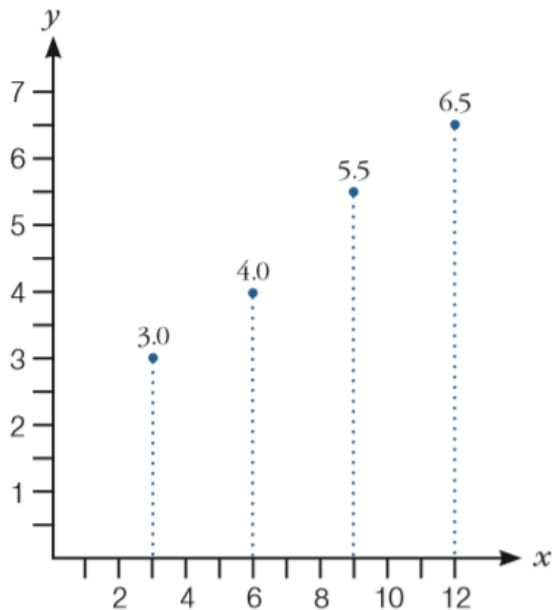
- All possible regression: 변수들의 모든 가능한 조합으로부터 최적의 모델을 찾아낸다. 탐색시간이 많이 드는 단점이 있음.
- Forward stepwise selection: 기여도가 높은 유의한 변수부터 하나씩 추가하는 방법. 탐색시간이 빠르다.
- Backward stepwise selection: 모든 변수를 포함한 상태에서 불필요한 변수를 제거해 나가는 방법. 중요변수가 제외될 가능성이 적음.

## 02-3 회귀모형의 결과분석 방법 (단순선형회귀 분석 예)

`lm(formula, data, ...)`: 선형회귀 모델을 생성하기 위한 함수

- `formula` : 반응변수 ~ 설명변수의 형태로 지정한 식
- `data` : 변수가 포함된 데이터 프레임

$X = \{3.0, 6.0, 9.0, 12.0\}$ ,  $Y = \{3.0, 4.0, 5.5, 6.5\}$



```
> x = c(3.0, 6.0, 9.0, 12.0)
> y = c(3.0, 4.0, 5.5, 6.5)
> m = lm(y ~ x)
> m
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x
1.75	0.40

$$y = 0.4x + 1.75$$

lm(formula, data=)로 구한 모델을 summary()로 요약하면

```
> summary(m)                # 모델의 상세 분석

Call:
lm(formula = y ~ x)

Residuals:
    1     2     3     4 
0.05 -0.15  0.15 -0.05 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.75000    0.19365   9.037  0.01202 *
x            0.40000    0.02357  16.971  0.00345 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1581 on 2 degrees of freedom
Multiple R-squared:  0.9931,    Adjusted R-squared:  0.9897 
F-statistic: 288 on 1 and 2 DF,  p-value: 0.003454
```



- **Residual**: 모델로 예측한  $y$ 값과 실제 데이터의  $y$ 값과의 차이를 의미한다. 패턴/추세가 보이면 안되며, residual plot으로 관측 가능하다.
- **Coefficient**: Estimate 컬럼: 절편과 각  $x$ 의 기울기 값이 출력된다. 또한 예측한 각 회귀계수에 대한 유의성을 나타낸다.
  - $t$ 값( $t$ 검정에 대한 통계량)은 독립변수( $x$ )와 종속변수( $y$ )간에 선형관계(관련성)가 존재하는 정도를 나타낸다.  $t$  값은 회귀계수 나누기 표준오차(표준편차)가 된다. 유의미한 결과가 나오려면  $t$  값이 커야 한다 (절대값이 2보다 커야 한다).  
(ex)  $1.75/0.19365=9.037$      $0.4/0.02357=16.971$
  - **유의수준(significance level)**:  $p$ -value로 표기되며, 관찰된 데이터의 검정 통계량이 귀무가설(영가설)을 지지하는 정도를 확률로 표현한 것. 유의수준 0.05로 설정한다면,  $p$ -값이 유의수준보다 작으므로 **귀무가설** (독립변수  $x$ 는 종속변수  $y$ 는 아무 관련이 없다)은 기각된다. 즉 두 변수는 관련이 있다라는 **대립 가설**이 받아들여진다.

Residuals:

1	2	3	4
0.05	-0.15	0.15	-0.05

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.75000	0.19365	9.037	0.01202 *
x	0.40000	0.02357	16.971	0.00345 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> plot(x, y)
> abline(m, col = 'red')
```

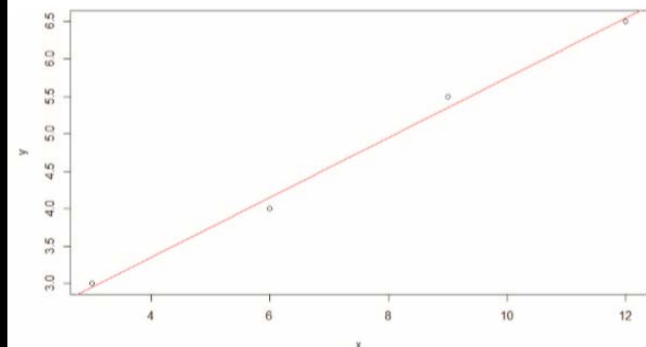


그림 7-4 lm 함수로 찾은 최적 모델

- **Multiple R-Squared**: 회귀모델의 설명력을 나타내는 결정계수로서 값이 1이면 실제 관측값들이 회귀선상에 정확히 일치함을 나타낸다. 만약 0.65이면 35%는 회귀식으로 설명할 수 없음을 의미한다. 단, 독립변수의 개수가 증가할수록 값이 증가하는 특징을 가지고 있다.
- **Adjusted R-Squared** : 독립 변수의 개수가 고려되어 보정된 R-Squared이다.
- **F-statistic**: 회귀식 전체의 유의성을 검정하는 값으로 가정 먼저 확인하여야 하는 값이다. “모든 회귀 계수가 0이다”라는 귀무가설( $H_0$ )의 기각여부를 검증하는 것이다. 즉 해당 p-value가 작은 값이면 최소 한 변수가 유의하다는 것을 알 수 있다. 어떤 변수가 유의한지는 coefficient table을 보면 된다.

Multiple R-squared: 0.9931,      Adjusted R-squared: 0.9897  
F-statistic: 288 on 1 and 2 DF,   p-value: 0.003454

## 02-4 회귀모형의 결과 분석 (다중선형회귀 분석 예)

```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-10.5932	-1.0690	0.2902	1.4272	3.3951

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.526667	0.374290	9.422	<2e-16 ***
youtube	0.045765	0.001395	32.809	<2e-16 ***
facebook	0.188530	0.008611	21.893	<2e-16 ***
newspaper	-0.001037	0.005871	-0.177	0.86

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 196 degrees of freedom  
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956  
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

```
Call:
```

```
lm(formula = sales ~ youtube + facebook, data = marketing)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-10.5572	-1.0502	0.2906	1.4049	3.3994

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.50532	0.35339	9.919	<2e-16 ***
youtube	0.04575	0.00139	32.909	<2e-16 ***
facebook	0.18799	0.00804	23.382	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.018 on 197 degrees of freedom  
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962  
F-statistic: 859.6 on 2 and 197 DF, p-value: < 2.2e-16

## 03 로지스틱 회귀분석

- 로지스틱 회귀분석: 분석 대상들이 두 집단 혹은 그 이상의 집단으로 나누어진 경우에 개별 관측치들이 어느 집단에 분류될 수 있는가를 분석하고 예측하는 모델
- 선형회귀 분석과 로지스틱 회귀분석의 비교

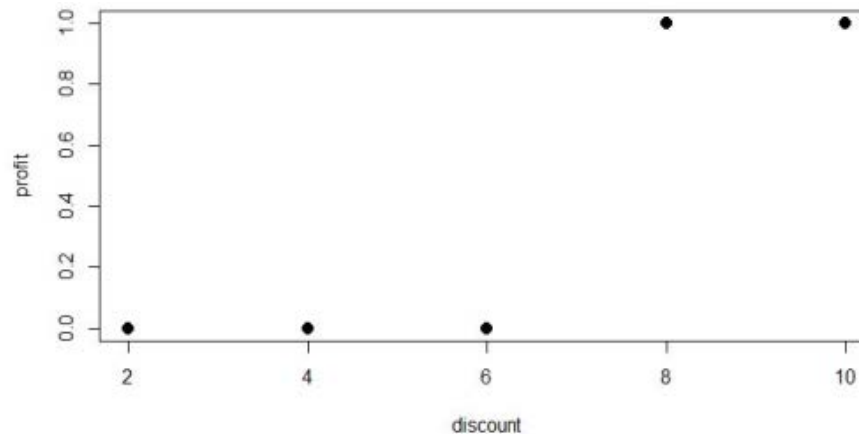
	일반선형회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	이산형 변수
모델 탐색 방법	최소제곱법	최대우도법(maximum likelihood method) 가중최소제곱법
모델 검정	F-test, t-test	$\chi^2$ test

- 로지스틱 회귀분석 과정
  - 각 집합에 속하는 확률의 추정치를 예측. 이진 분류의 경우에는 집단 1에 속하는 확률  $P(Y=1)$ 의 추정치를 얻는다.
  - 확률값  $\rightarrow$  분류 기준값 (cut-off) 적용  $\rightarrow$  특정 집단으로 분류  
예)  $P(Y=1) \geq 0.5 \rightarrow$  집단 1로 분류  
 $P(Y=1) < 0.5 \rightarrow$  집단 0으로 분류

## 03-1 (이항) 로지스틱 회귀분석

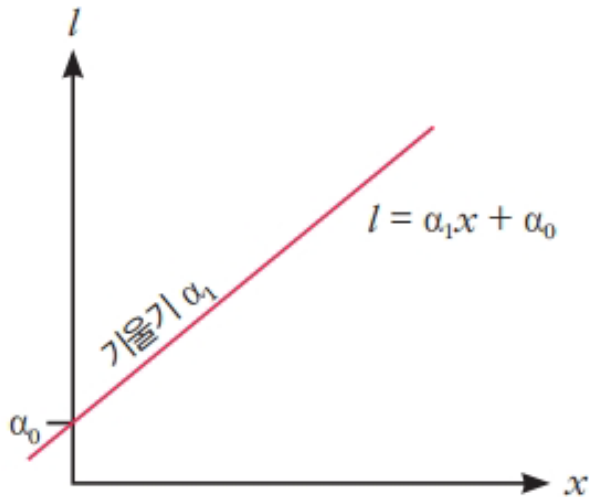
- 종속(반응) 변수가 두 가지 값만 가지는 경우
  - 의사가 검사 결과를 보고 환자와 정상을 구분
  - 농장에서 따온 과일을 상품과 하품으로 구분하는 경우
- (예제) 할인율에 따른 이익을 산정하는 판매 데이터
  - 상품을 팔면서 데이터를 수집.
  - 순이익은 5만원 미만이면 0, 넘으면 1로 기록

discount	profit
2	0
4	0
6	0
8	1
10	1



## 03-2 로지스틱 회귀 – 원리 분석

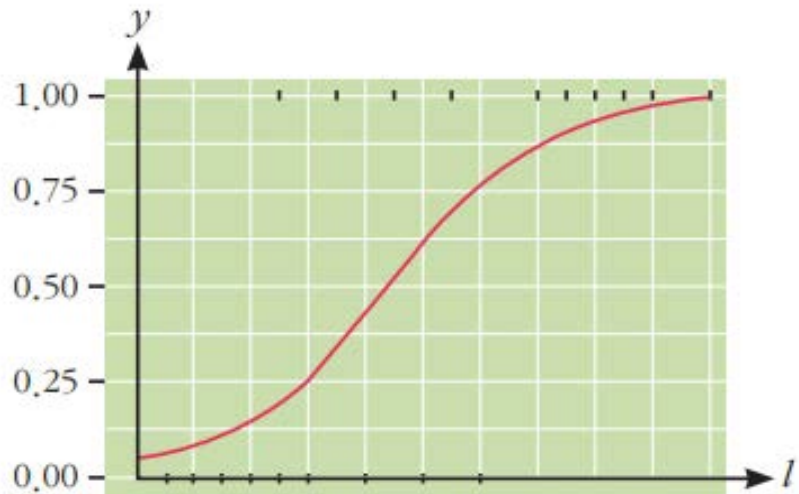
- 종속 변수가 두 가지 값만 가지는 경우의 회귀  
(참/거짓, 성공/실패, 환자/정상, 사망/생존, 승리/패배 등)
- 원리
  - 독립 변수를  $x$ , 종속 변수를  $l$ 로 표기 (다음 그림에서 가로축은  $x$ , 세로축은  $l$ 을 나타냄)



(a) 선형 회귀 함수

$$l = \alpha_1 x + \alpha_0$$

식 (1)



(b) 로짓 함수

$$y = \frac{1}{1 + e^{-l}}$$

식 (2)

## 03-2 로지스틱 회귀

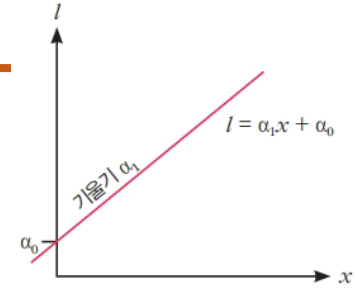
### • 원리

- $$l = \alpha_1 x + \alpha_0 \quad \text{식 (1)}$$

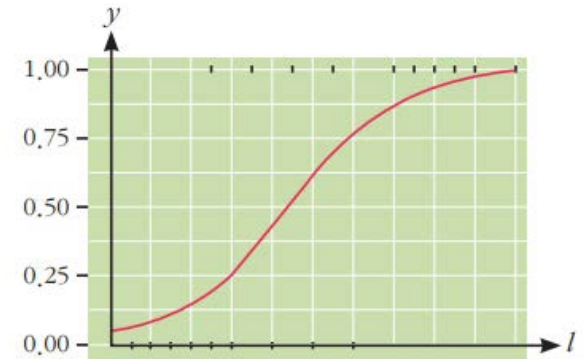
- 반응 변수  $l$ 의 범위는  $[-\infty, \infty]$ 이므로 로지스틱 회귀를 모델링할 수 없음
- 해결책: 로짓 함수 logit function라 부르는 식 (2)를 추가로 사용 → 범위를  $[0,1]$ 로 축소

- $$y = \frac{1}{1 + e^{-l}} \quad \text{식 (2)}$$

- 가로축은  $l$ , 세로축은  $y$ 를 나타내며  $y$ 는  $[0,1]$  사이로 축소되었음에 주목.
- $l$ 을 잠복 latent 변수 또는 은닉 hidden 변수라 부름



(a) 선형 회귀 함수



(b) 로짓 함수

### • 일반화 선형 회귀는 두 단계 변환을 사용

- 일반화 선형 회귀에는 로지스틱 회귀뿐 아니라 지수 회귀, 포와송 회귀 등이 있음
- 로지스틱 회귀는 식 (1)과 식 (2)를 사용
- 식 (2)와 같은 함수를 링크 함수라 부름 (로지스틱 회귀는 링크 함수로 로짓 함수를 사용하는 셈)

$$L = \alpha_1 x + \alpha_0$$

회귀식의 장점은 그대로 유지하되 종속변수  $L$ 를 범주가 아니라 (범주1이 될)확률로 두고 식을 세운다.  
우변은 그대로 두고 좌변만 확률로 바꾸면

$$P(L=1/x) = \alpha_1 x + \alpha_0$$

좌변의 범위는 0~1 사이이다. 하지만 우변은 음의 무한대에서 양의 무한대 범위를 가지기 때문에 식이 성립하지 않는 경우가 존재할 수 있다. 여기서 식을 한번 더 바꿔서, 좌변을 승산(odds)으로 설정해 본다.

$$P/1-P = \alpha_1 x + \alpha_0$$

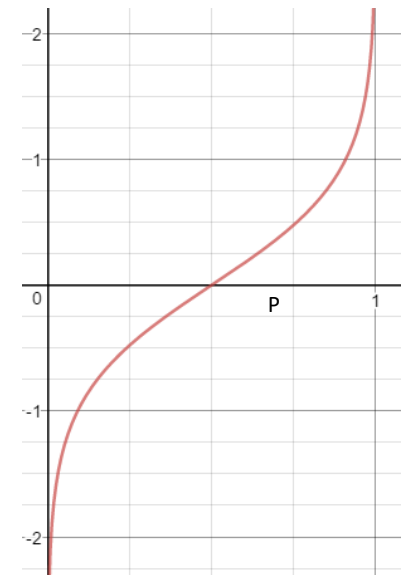
좌변(승산)의 범위는 0에서 무한대의 범위를 갖는다. 하지만 우변(회귀식)은 그대로 음의 무한대에서 양의 무한대 범위입니다. 다시 여기에서, 좌변(승산)에 로그를 취하여 본다.

$$\log(P/1-P) = \alpha_1 x + \alpha_0$$

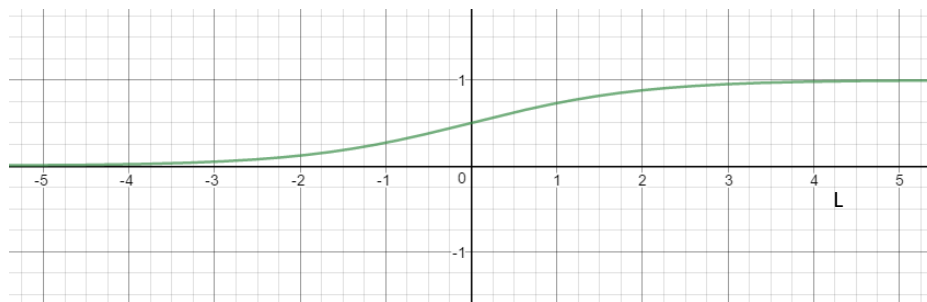
로그 승산의 범위 또한 우변처럼 음의 무한대에서 양의 무한대가 된다.  
이제야 비로소 좌변(승산)이 우변(회귀식)의 범위와 일치하게 된다.  $P$ 에 대하여 정리하면,

$$(P/1-P) = e^{\alpha_1 x + \alpha_0}$$

$$P = e^{\alpha_1 x + \alpha_0} / 1 + e^{\alpha_1 x + \alpha_0} = 1 / 1 + e^{-(\alpha_1 x + \alpha_0)}$$



log(P/1-P)의 그래프



$1/1 + e^{-L}$ 의 로짓함수 그래프



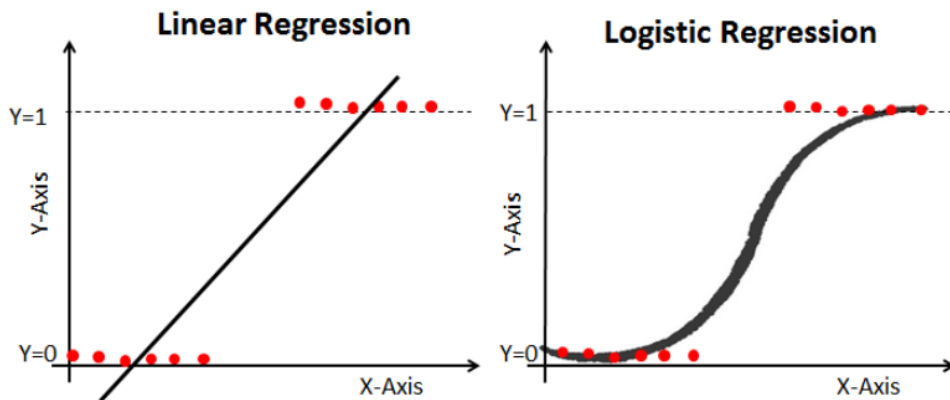
## 03-3 로지스틱 회귀 – glm() 함수사용

- R에서 일반화 선형 모델은 glm 함수로 수행

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

- formula: 모형식, 반응변수 ~ 설명변수들
- data: 모형식을 적용할 데이터프레임
- family: 오차 분포와 링크 함수(link function), 로지스틱 회귀 모형의 경우 family = "binomial"을 지정합니다.

단순/다중 선형회귀분석에서는 종속변수  $y$ 가 연속형으로 가정되었다.  
로지스틱 회귀분석은 종속변수가 범주형으로 0 또는 1인 경우 사용하는 회귀분석이다.



# 03-4 로지스틱 회귀 분석 - 예제

- UCLA admission 데이터 (400 obs. Of 4 variables)

```
> m = glm(admit~., data = ucla, family = binomial)
> coef(m)
(Intercept)      gre      gpa      rank
-3.44954840  0.00229396  0.77701357 -0.56003139
```

$$y = \frac{1}{1 + e^{-l}} \quad \text{식 (2)}$$

$$P = 1 / (1 + e^{-(0.002gre + 0.77gpa - 0.56rank - 3.44)})$$

```
> s = data.frame(gre = c(376), gpa = c(3.6), rank = c(3))
> predict(m, newdata = s, type = 'response')
1
0.1869631
```

합격 확률은 18.7%

admit	gre	gpa	rank
0	380	3.61	3
1	660	3.67	3
1	800	4	1
1	640	3.19	4
0	520	2.93	4
1	760	3	2
1	560	2.98	1
0	400	3.08	2
1	540	3.39	3
0	700	3.92	2
0	800	4	4
0	440	3.22	1
1	760	4	1
0	700	3.08	2

- admit : 불합격은 0, 합격은 1
- gre : 미국 대학원 수학능력시험인 gre의 점수
- gpa : 학부 성적(평균 학점)
- rank : 출신 대학 순위, {1, 2, 3, 4}의 4개 값

# 03-5 로지스틱 회귀 – 결과 분석

지원자의 학교 등급을 나타내는 rank변수는 범주형이므로 factor처리 해 줍니다.

```
> summary(model)
```

```
call:
glm(formula = admit ~ gre + gpa + rank, family = "binomial",
    data = train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6154	-0.7872	-0.5325	0.8785	2.3242

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.790221	1.750305	-2.165	0.030352	*
gre	0.003483	0.001718	2.027	0.042640	*
gpa	0.534007	0.519132	1.029	0.303642	
rank2	-0.707194	0.476566	-1.484	0.137826	
rank3	-1.803655	0.531352	-3.394	0.000688	***
rank4	-1.959514	0.629189	-3.114	0.001844	**

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 237.18 on 199 degrees of freedom  
Residual deviance: 205.76 on 194 degrees of freedom  
AIC: 217.76

Number of Fisher Scoring iterations: 4

deviance residuals: 모델 fitting이 잘 되었는지에 대한 척도를 나타냄.

회귀계수와 그것들의 표준편차, z-statistics(wals's z-statistics), p-value를 나타낸다.

Null deviance와 Residual deviance는 각각 절편모형과 제언 모형의 완전모형으로부터의 이탈도를 나타내며 값이 작을수록 해당 모형이 자료에 적합함을 나타낸다.

이탈도에 기초한 구체적인 검정은 카이제곱 분포를 이용한다.

# 03-5 로지스틱 회귀

```
> summary(model)
```

```
Call:
glm(formula = admit ~ gre + gpa + rank, family = "binomial",
    data = train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6154	-0.7872	-0.5325	0.8785	2.3242

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.790221	1.750305	-2.165	0.030352	*
gre	0.003483	0.001718	2.027	0.042640	*
gpa	0.534007	0.519132	1.029	0.303642	
rank2	-0.707194	0.476566	-1.484	0.137826	
rank3	-1.803655	0.531352	-3.394	0.000688	***
rank4	-1.959514	0.629189	-3.114	0.001844	**

---  
signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 237.18 on 199 degrees of freedom  
Residual deviance: 205.76 on 194 degrees of freedom  
AIC: 217.76

Number of Fisher Scoring iterations: 4

회귀분석

로지스틱 회귀분석

모형에 대한 검정

F검정

카이제곱

계수에 대한 검정

T검정

Wald 통계량

설명력

R<sup>2</sup>

Cox and Snell R<sup>2</sup> 또는 Nagelkerke R<sup>2</sup>

## 04-1 교차 분석 (cross tabulation analysis)

- 두 범주형 변수 간에 연관관계(association)가 있는지의 여부를 판단하고자 하는 경우 교차표를 작성하여 변수들 간의 관계를 분석한다.
- 이를 교차분석 혹은  $\chi^2$ (chi-square) 검정(test)이라고 한다. 교차분석은 두 변수의 빈도표를 교차 시킨다는 의미로 해석되며, 교차분석에 사용되는 검정 통계량이  $\chi^2$ -분포를 따르므로  $\chi^2$ -검정으로 부른다.
- 교차표(cross tabulation), 분할표(contingency table): 각 범주형 변수에 대한 빈도표를 행과 열로 결합시켜 놓은 형태이다. 일반적으로 행에는 설명(독립)변수에 해당하는 변수를 할당하고, 열에는 반응(종속)변수를 할당한다.
- 2×2 교차표 :

성별	전공			Total
	A 전공	B 전공	C 전공	
남자	75	46	23	144
여자	30	32	24	86
Total	105	78	47	230

## 04-2 $\chi^2$ - 검정에 의한 독립성 검정

- 교차표 : 성별과 전공 선택이 서로 관계가 있을까?

성별	전공			Total
	A 전공	B 전공	C 전공	
남자	75	46	23	144
여자	30	32	24	86
Total	105	78	47	230

- 가설:

- 귀무가설: 두 변수는 독립이다. 남녀별 전공선택의 차이는 없다
- 대립가설: 서로 독립이 아니다. 남녀별 전공선택의 차이는 있다.

- 검정통계량:

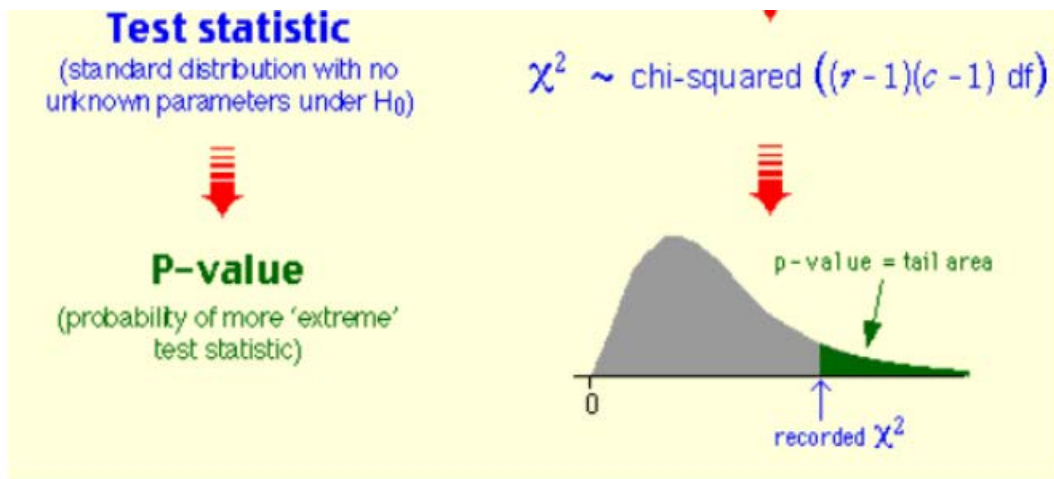
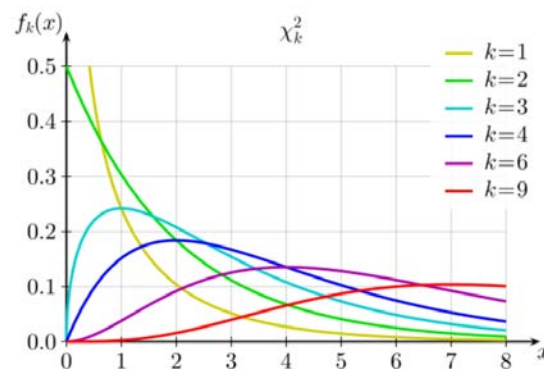
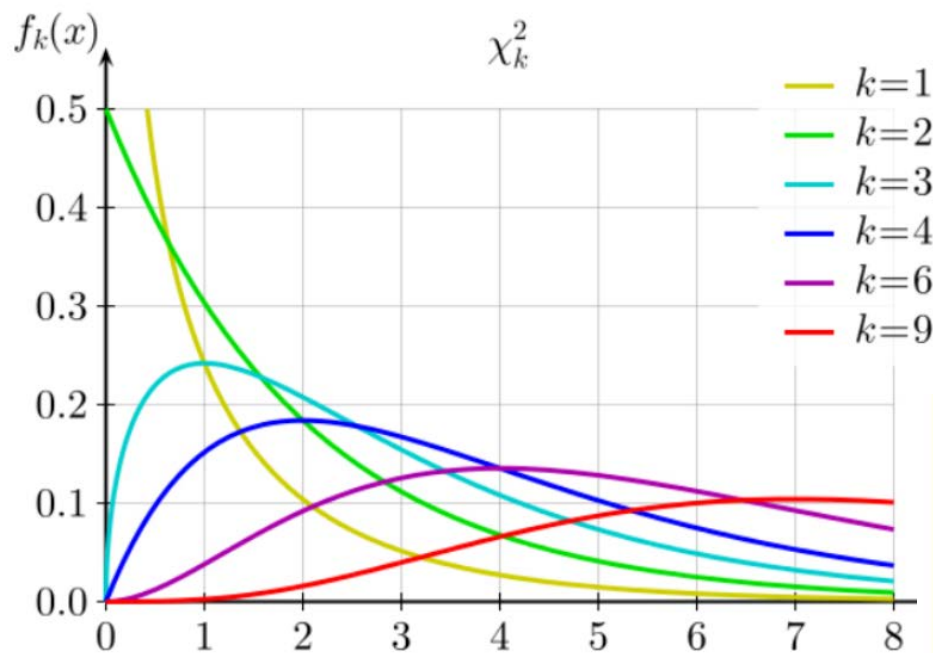
$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- 두 변수가 서로 독립이라면  $P(AB)=P(A)P(B)$ 가 성립한다. Ex)  $P(\text{남자} \cap \text{A전공}) = P(\text{남자})P(\text{A전공})$
- 기대 빈도( $E_{ij}$ ): 두 변수가 독립이라는 가정 하의 i-행 j-열 셀의 빈도 =  $\frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$
- 관측 빈도( $O_{ij}$ ): 표본으로부터 관측된 빈도
- 의미: 만약 두 변수가 독립이라면 ( $O_{ij}=E_{ij}$ )이고 T 값은 0일 것이다. 즉 T가 0에 가까우면 두 변수는 관계가 없다고 결론 내릴 수 있다. 또한 이 통계량은  $\chi^2$  ( $df=(R-1)(C-1)$ )에 근사함이 밝혀져 있다. 따라서 T 값이 커지면 두 변수는 관계가 있다고 결론 지을 수 있다.

✓ 여기에서, R: 행의 수, C: 열의 수, df(degree of freedom): 자유도

## 04-2 $\chi^2$ 분포

- $\chi^2$  분포는  $k$ 개의 서로 독립적인 표준정규 확률변수를 각각 제공한 후, 이들을 합하여 얻어지는 분포이다. 이 때  $k$ 를 자유도라고 하며, 매개 변수가 된다.
- 항상 양의 값을 가지며 비대칭(오른쪽 긴 꼬리)적인 분포 모양을 갖는다.
- 자유도에 따라 분포의 모양이 변하며, 자유도가 커질수록 정규분포에 가까워진다.



# 04-3 $\chi^2$ - 검정에 의한 독립성 검정 예제

기대빈도( $E_{11}$ )= $105 \times 144 / 230 = 65.7$

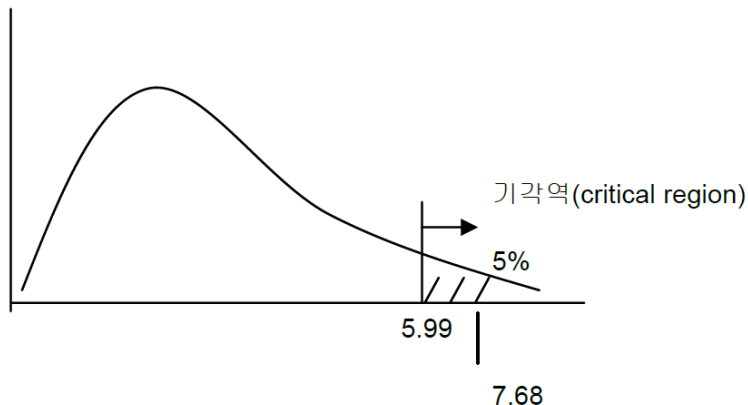
$E_{23} = 86 \times 47 / 230 = 17.6$

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

성별	전공			Total
	A 전공	B 전공	C 전공	
남자	75	46	23	144
여자	30	32	24	86
Total	105	78	47	230

$$T = (75 - 65.7)^2 / 65.7 + \dots + (24 - 17.6)^2 / 17.6 = 7.68$$

자유도  $(R-1)(C-1) = (2-1)(3-1) = 2$



계산된 검정 통계량이 7.68이므로 기각역에 속하여 귀무가설이 기각되고 그 결과 성별과 전공선택에는 관계가 있다고 말할 수 있다.



# 04-4 R code 예제(1)

예제: Smoke (smoking habit)와 Exer(exercise level) 사이에 연관성 검사

```
> library(MASS)          # load the MASS package
> tbl = table(survey$Smoke, survey$Exer)
> tbl                     # the contingency table
```

	Freq	None	Some
Heavy	7	1	3
Never	87	18	84
Occas	12	3	4
Regul	9	1	7

```
> chisq.test(tbl)
```

Pearson's Chi-squared test

```
data:  table(survey$Smoke, survey$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828
```

Warning message:

```
In chisq.test(table(survey$Smoke, survey$Exer)) :
  Chi-squared approximation may be incorrect
```

```
> ctbl = cbind(tbl[, "Freq"], tbl[, "None"] + tbl[, "Some"])
> ctbl
```

	[,1]	[,2]
Heavy	7	4
Never	87	102
Occas	12	7
Regul	9	8

```
> chisq.test(ctbl)
```

Pearson's Chi-squared test

```
data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571
```

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl, and save it in a new table named ctbl. Then we apply the chisq.test function against ctbl instead.

## 04-4 R code 예제(2)

예제: left (이직여부)와 salary(연봉수준) 변수  
사이에 연관성 검사

```
library(gmodels)

HR = read.csv("C:/R/HR_comma_sep.csv")
HR$salary = factor(HR$salary ,levels = c('low','medium','high'))

CrossTable(HR$left,HR$salary,
            prop.r = FALSE,prop.c = FALSE,prop.t = FALSE,
            chisq = TRUE,expected = TRUE)
```

- ✓ prop.r, prop.c, prop.t: 각 셀의 비율표시에 대한 옵션
- ✓ Chisq=TRUE: 카이제곱 독립성 검정실행
- ✓ Expected=TRUE: 기대빈도 표시

Cell Contents

-----
N
Expected N
Chi-square contribution
-----

Total Observations in Table: 14999

	HR\$salary			
HR\$left	low	medium	high	Row Total
-----				
0	5144	5129	1155	11428
	5574.188	4911.320	942.492	
	33.200	9.648	47.915	
-----				
1	2172	1317	82	3571
	1741.812	1534.680	294.508	
	106.247	30.876	153.339	
-----				
Column Total	7316	6446	1237	14999
-----				

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 381.225      d.f. = 2      p = 1.652087e-83

- 회귀분석

- ✓ 선형회귀분석
  - ✓ 단순선형회귀분석
  - ✓ 다중선형회귀분석
- ✓ 일반화 선형회귀분석
  - ✓ 로지스틱 회귀분석
- ✓ 결과 분석 방법

- 교차분석 ( $\chi^2$  test): 두 범주형 변수 간의 연관관계 분석