선형회귀분석

회귀분석 이란?

영국의 유전학자 프란시스 골턴(F. Galton, 1889)에 의해 소개된 것으로 사람의 키에 관한 유전연구를 통해 회귀(回歸, regression)라는 생각을 발전시킴

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the mid-parent in inches	Height of the adult child														Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

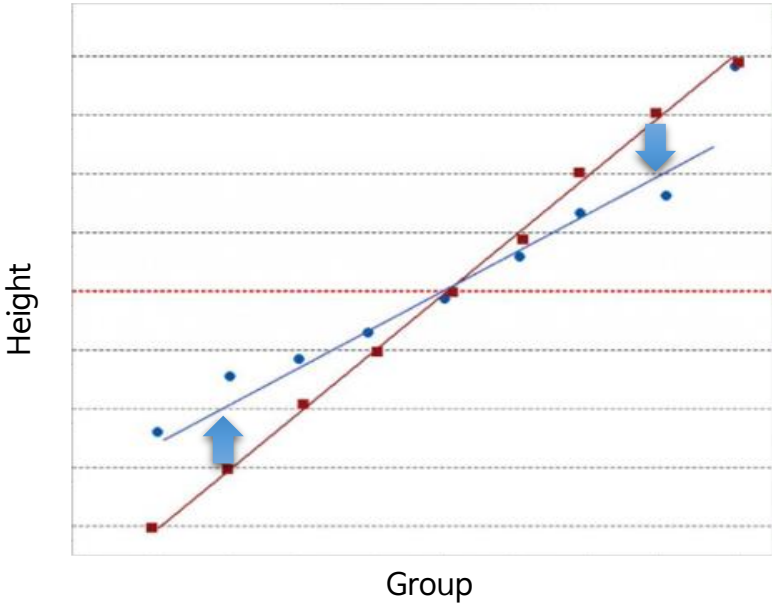
Source: Galton (1886a).

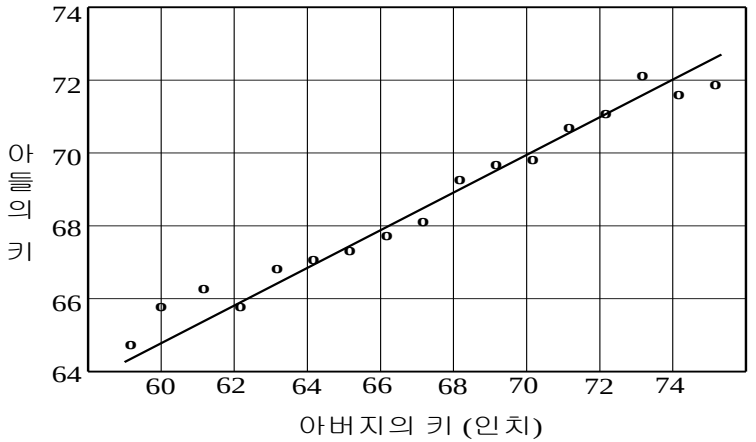
Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties); he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

회귀분석 명칭의 유래

Galton의 연구

- 아버지의 키 그룹별로 대응되는 아들 그룹의 평균 키를 표시하고, 이러한 관계를 잘 나타낼 수 있는 직선을 표현
- K. Pearson(1903)은 아버지의 키(x)와 아들의 키(y)에 관한 기록을 수집하여 일차함수 관계식을 얻음





K. Pearson (1903) :

$$Y = 33.73 + 0.516X$$

부모의 키가 크면 자녀의 키도 크지만, 자녀의 키는 평균키로 **회귀함**

Regression 용어 사용

회귀분석 이란?

- 자연현상이나 사회현상들은 관련된 여러 변수들의 상호작용에 의해 발생

← 과학은 현상들을 이해하고 규명하고자 하는 것이고 통계학은 방법을 제시(통계학은 과학의 문법이다.)

- 현상들이 어떤 변수들의 관계에 의해 나타날 때 그 관계를 수학적으로 설명하기 위해서 사용되는 방법

회귀모형 예시

$$\text{자녀의 키} = \beta_0 + \beta_1 * \text{부모의 키} + \beta_2 * \text{환경적인 요인} + \varepsilon$$

종속변수 (y)

- 반응(Response) 변수
- 독립변수들에 의해 설명되는 변수

독립변수 (x_i)

- 설명(Explanatory) 변수
- 설명에 이용되는 변수

회귀분석 이란?

종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 통계적 기법

회귀모형 예시

자녀의 키 = β_0 + β_1 * 부모의 키 + β_2 * 환경적인 요인 + ε

절편항 β_0

- 독립변수가 모두 0 일 때의 종속변수의 값
- 독립변수에 영향을 받지 않는 값

기울기 β_i

- 독립변수(x_i) 가 1단위 증가할 때마다 증가하는 종속변수(y)의 양

오차항 ε

- 회귀식과 실제 종속변수와의 차이

변수들 간의 관계를 나타내는 수학적 모형

결정적 모형(Deterministic Model)

변수들 간의 관계가 오차(error)가 존재하지 않는 정확한 수학적 함수관계로 설정된 모형

물리 법칙 등

통계적 모형(Statistical Model)

변수들 간의 관계에 오차를 허용하는 모델

① 변수들의 값을 관측하는 경우 측정오차 발생

② 변수들 간의 관계가 정확하게 알려져 있지 않거나 또는 알려져 있다 하더라도 복잡한 형태로 주어지는 경우 근사적 모형을 사용한다.

③ 모든 변수들을 모형에 포함시켜 분석하는 것은 어려우므로 특정 변수 이 외의 영향 인자들은 오차항에 포함시킨다.

회귀분석을 하는 목적

회귀분석을 하는 목적은 다음 3가지 정도로 정리해 볼 수 있다.

① 변수들 간에 성립하는 정확한 회귀모형의 구축(model building)

→ 모형의 구축에서는 독립변수들의 선택과 회귀식의 형태가 관심의 대상

② 모형에 포함된 모수들의 추정(parameter estimation)

→ 모수의 추정에서는 추정 방법과 추정값의 통계적 유의성 여부가 중요

③ 추정된 모형을 이용한 예측(prediction)

→ 예측을 목적으로 하는 경우에는 추정된 회귀모형의 예측정확도를 조사할 필요가 있다.

단순선형회귀모형

종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 통계적 기법

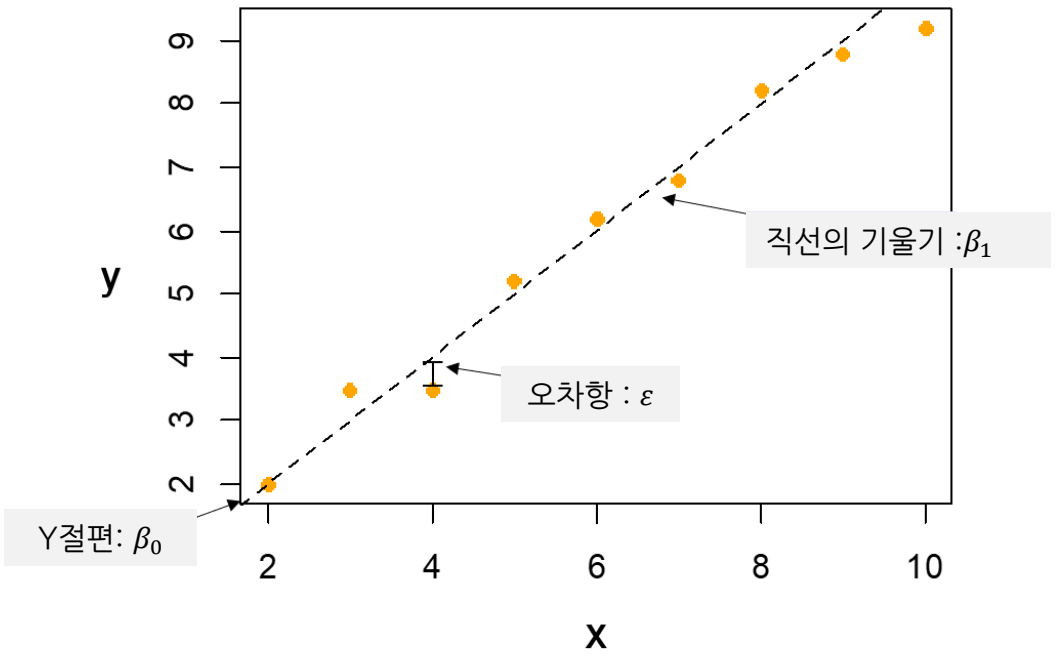
단순선형회귀모형

- **정의**
  - ✓ 독립변수가 한 개인 선형모형의 분석 기법
  - ✓ 회귀식이 모수들의 일차식으로 표현됨

- **기본식**

$$y = \beta_0 + \beta_1 x + \varepsilon$$

β_0 : 절편(intercept), β_1 : 기울기(slope)



단순선형회귀모형

종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 통계적 기법

단순선형회귀모형

$$y = \beta_0 + \beta_1 x + \varepsilon$$

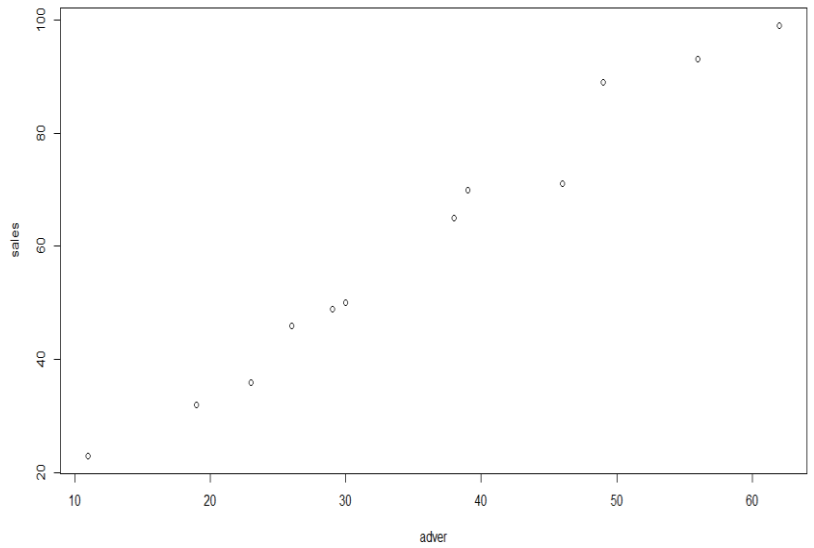
- 회귀계수 β_0와 β_1 을 추정하기 위해서는 종속변수 Y와 독립변수 X의 관측값의 표본이 필요
- 이들 관측값들은 일반적으로 모두 일직선 상에 위치하지 않는다.
  - ✓ Y와 X가 선형의 관계를 가지고 있다 하더라도 관측값에는 측정오차가 있을 수 있고,
  - ✓ 실제 Y와 X 가 정확하게 선형관계를 형성하지 않을 수도 있기 때문
- ε_i는 평균이 0이고 분산이 σ^2인 오차를 나타내는 확률변수로 관측값 Y_i가 모집단 회귀식에서 ε_i 만큼 떨어져 있다는 것을 의미

단순선형회귀모형

R CODE

```
## 예제 데이터
# 유사한 제품을 생산하는 12개 기업에 대해
# 1년 광고비(독립변수, x)와 매출액(종속변수, y)

x <- c(11, 19, 23, 26, 56, 62, 29, 30, 38, 39, 46, 49)
y <- c(23, 32, 36, 46, 93, 99, 49, 50, 65, 70, 71, 89)

plot(y~x, xlab="adver", ylab="sales")
```

R OUTPUT

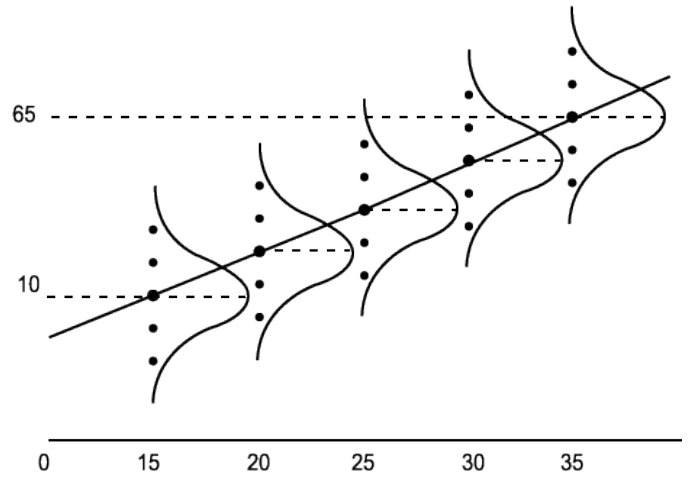


회귀분석의 기본 가정

회귀분석을 위해서는 기본적인 가정이 필요하다.

- **독립성**
  - ✓ 오차항들은 서로 **독립적** 이어야 함
- **정규성**
  - ✓ 오차항의 분포는 **평균이 0**인 **정규분포**임
- **등분산성**
  - ✓ 오차항의 **분산이 일정**해야 함

- **오차항 (ε)**
  - ✓ 회귀선과 실제 값의 차이를 의미하며 이러한 오차들이 모여 분포를 이루게 됨



회귀분석모형의 종류

종속변수가 독립변수들에 의해 어떻게 설명되는지를 분석하는 통계적 기법

명칭	모형	특징
단순 선형 회귀분석	$y = \beta_0 + \beta_1 \boldsymbol{x_1} + \epsilon$	독립변수가 한 개인 경우
다중 회귀분석	$y = \beta_0 + \beta_1 \boldsymbol{x_1} + \ldots + \beta_2 \boldsymbol{x_p} + \epsilon$	독립변수가 두 개 이상인 경우
다항 회귀분석	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{12} \boldsymbol{x_1}\, \boldsymbol{x_2} + \gamma_1 \boldsymbol{x_1}^2 + \ldots + \epsilon$	독립변수들 사이의 교차 영향이나 제곱 영향을 고려하는 경우
다변량 회귀분석	$\boldsymbol{y_1} = \beta_{10} + \beta_{11} \boldsymbol{x_1} + \ldots + \beta_{1p}\, \boldsymbol{x_p} + \epsilon$ $\boldsymbol{y_2} = \beta_{20} + \beta_{21} \boldsymbol{x_1} + \ldots + \beta_{2p}\, \boldsymbol{x_p} + \epsilon$	종속변수들 간의 상관관계가 있을 경우

회귀 계수의 추정

회귀분석에서 회귀 계수는 최소제곱적합법에 의해 추정한다.

최소제곱적합법 (LSE, Least Square Estimation)

- **개념**
  - ✓ 회귀계수 β는 미지의 모수이므로 표본 (X_i, Y_i) 을 이용하여 β를 추정한다는 것은 미지의 직선식을 적합하는 것
  - ✓ 근사적으로 구한 추정 값과 실제 값의 차(오차)의 제곱합(S)을 최소로 하는 β를 추정
  - ✓ 모든 점에서 잔차값을 최소화할 수는 없고 잔차의 크기를 전체적으로 작게 하는 방법을 선택(최소제곱법)
- **최소 제곱 방정식**

구분	단순 회귀분석 모형	다중 회귀분석 모형
회귀식	$y = \beta_0 + \beta_1 X$	$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
오차 제곱 합 (S)	$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$	$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$
최소 제곱 방정식	$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$	$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p$

회귀모형을 이용한 예측 방법

회귀분석에서 회귀 계수는 최소제곱적합법에 의해 추정한다.

회귀계수 해석 및 예측 예시

- 광고비(x)와 매출액(y)을 이용한 선형회귀분석 시행

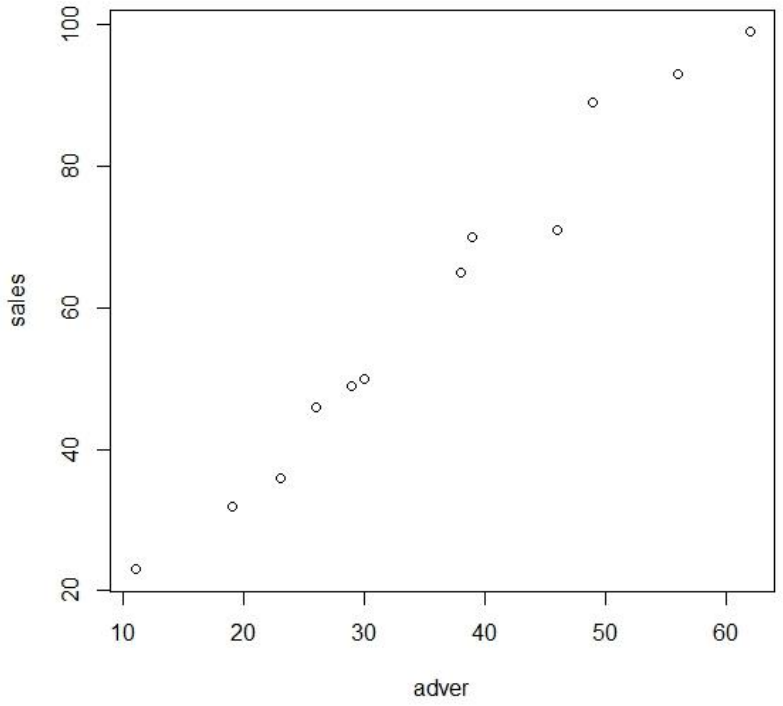
$$\hat{y} = 3 + 1.5x$$

✓ $\widehat{\beta_0} = 3$: 광고비가 0 일 때 매출 추정값

✓ $\widehat{\beta_1} = 1.5$: 광고비가 1 증가할 때 매출액은 1.5 증가

✓ 예측 : 광고비를 10만큼 투자하고자 할 때 예상 매출액은?

: 3 + 1.5×10 = ***18***

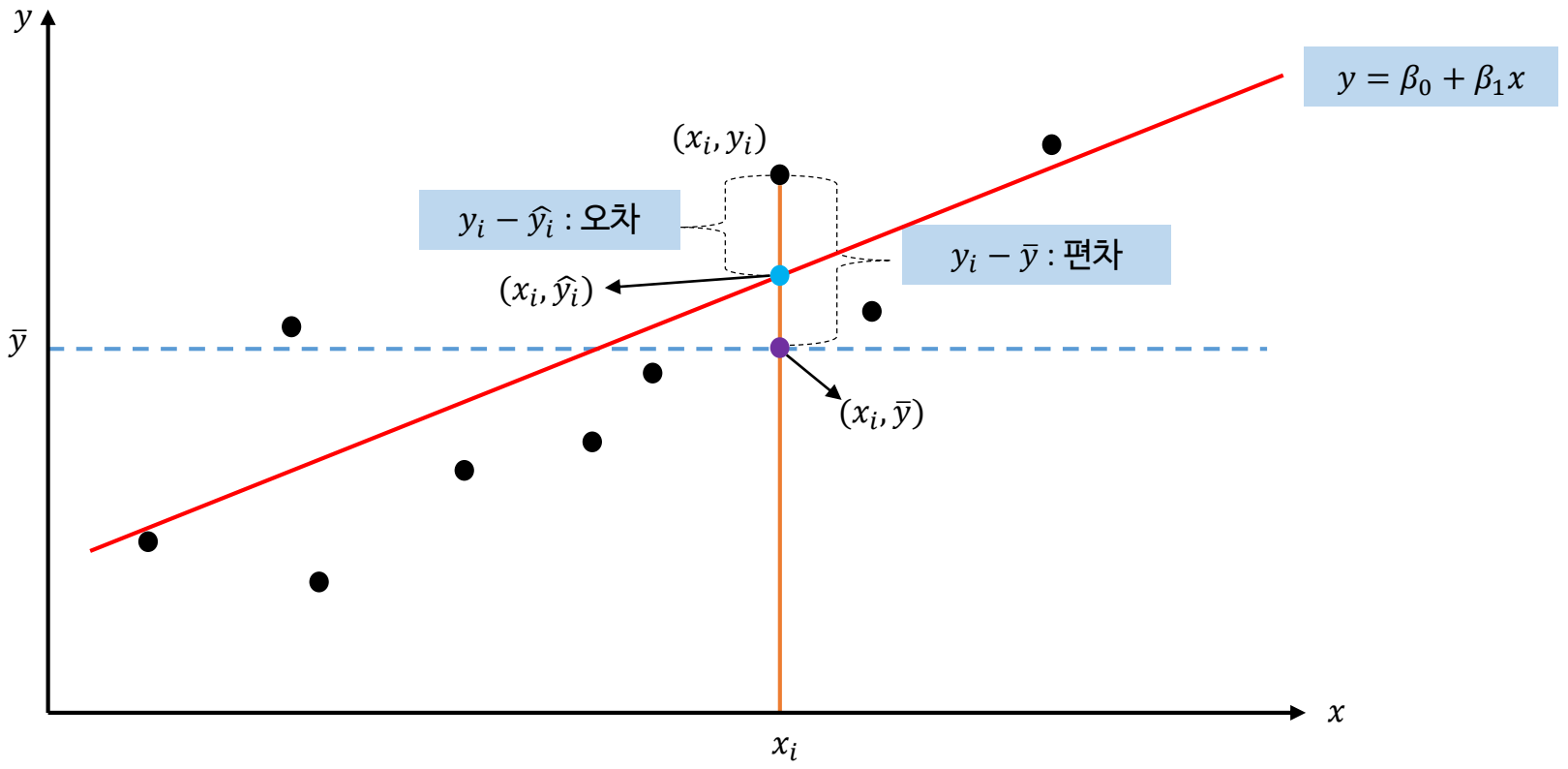


광고비 10 증가하면 매출액은 15만큼 증가될 것이라 예측

회귀모형의 해석

추정된 회귀식은 결정계수, 분산분석표를 활용하여 해석한다.

편차와 오차



회귀모형의 해석

추정된 회귀식은 결정계수, 분산분석표를 활용하여 해석한다.

제곱합(Sum of Square)의 분할

$$\sum_{i=1}^{n}(y_i - \bar{y_i})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y_i})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

총 제곱합 (TSS)
Total sum of squares

- 종속변수 값의 총 변동
- 종속변수의 산포를 나타내는 값

회귀 제곱합 (SSR)
Sum of squares for regression

- 총 변동 중에서 회귀식에 의해 설명되는 변동

오차 제곱합 (SSE)
Sum of squares for residual

- 총 변동 중에서 오차에 의해서 생기는 변동

회귀모형의 해석

추정된 회귀식은 결정계수, 분산분석표를 활용하여 해석한다.

결정계수 (R-Square)

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$$

결정계수

- 회귀식의 적합도(Goodness of fit)를 측정하는 한가지 척도
- 전체 제곱합 중에서 회귀 제곱합이 차지하는 비율
- 종속변수의 총 변동을 회귀식이 얼마나 설명하는가를 나타냄

회귀모형의 해석

추정된 회귀식은 결정계수, 분산분석표를 활용하여 해석한다.

분산분석표(ANOVA Table)

요인	제곱합	자유도	평균제곱	분산비
회귀	SSR	p	MSR=SSR/p	F_0=MSR/MSE (p-value)
오차	SSE	n-p-1	MSE=SSE/(n-p-1)	
전체	TSS	n-1		

- 분산비 F값의 유의확률(p-value)을 통해 회귀식의 유의성 검정 가능
  - ✓ $H_0 : \beta_1 = 0$
  - ✓ $H_1 : \beta_1 \neq 0$

  - p-value < 0.05이면 귀무가설 기각

모형의 해석

② Root MSE = 3.834 R-Sq = 97.9% R-Sq(adj) = 97.6%

Analysis of Variance

①

Source	DF	SS	MS	F	P
Model	1	6695.3	6695.3	455.54	0.000
Error	10	147.0	14.7		
Total	11	6842.2			

- 회귀식의 유의성(F-검정, ①)
  - 유의수준: 0.05일 경우, p-value가 유의수준보다 작으므로 귀무가설을 기각함(대립가설 채택)

- 회귀식의 설명력(결정계수, ②)
  - 종속변수의 변동 중 97.9%가 회귀식에 의해 설명됨

회귀모형의 해석

추정된 회귀식은 결정계수, 분산분석표를 활용하여 해석한다.

F-검정

- F-비가 자유도가 1과 (n-2)인 F 분포에서의 상위 α 분위수보다 크면, 즉,

$$F_0 > F(\alpha; 1, n-2)$$

- 이면 가설 $H_0 : \beta_1 = 0$을 유의수준 α 하에서 기각
- 가설 $H_0 : \beta_1 = 0$ 가 성립하는 경우,
- 종속변수 Y와 독립변수 X 가 서로 관련성이 없다는 것을 의미
- 이 가설을 기각하는 경우에는 독립변수 X가 종속변수 Y에 대해 통계적으로 유의한 영향을 준다고 해석

모형 해석의 유의할 점

- $H_1 : \beta_1 \neq 0$ 을 채택하는 결론은 단지 두 변수간에 선형관계가 존재한다는 것만을 나타낼 뿐 추정된 회귀직선이 유용하다는 직접적인 의미가 아니다.
  - ✓ $\beta_1 = 0$ 인 경우 회귀직선의 적합도가 높을 가능성은 있지만 다른 중요한 독립변수가 분석에 포함되지 않았다든지
  - ✓ 또는 잘못된 모형 설정으로 인하여 적합도가 낮을 수 있기 때문이다.
- 회귀모형의 유용성에 대한 결론은 F-검정의 결과, 결정계수와 잔차표준오차 등을 종합적으로 참조해야 한다.

회귀모형의 해석 실습

R CODE

```
# 유사한 제품을 생산하는 12개 기업에 대해
# 1년 광고비(독립변수, x)와 매출액(종속변수, y)

# (1) 분산분석표를 작성하고 F-검정을 실시

adsales.lm <- lm(y ~ x)
anova(adsales.lm)
```

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1 6695.3  6695.3  455.54 1.136e-09 ***
Residuals 10  147.0    14.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

결과 해석

- SSE에 대한 정보는

  'Residuals' 행에 나와 있다. 자유도는 10이고 SSE는 147.0 이다.

- SSR에 대한 정보는

  'X' 행에 나와 있다. 자유도는 1이고 SSR은 6695.3 이다.

- $F_0 = 455.54$
- F(0.05, 1, 10) = 4.96 보다 크므로 유의수준 α=0.05 에서 H_0 : $\beta_1 = 0$ 을 기각
- F-비에 대한 해석은, 설명된 변동의 크기가 설명되지 않은 변동의 크기의 약 455배가 된다는 것이다.
- 즉, 독립변수가 종속변수의 변동을 충분히 설명하므로 회귀식에 독립변수를 포함시키는 것이 타당하다는 결론을 도출

회귀모형의 해석 실습

R CODE

```
# 유사한 제품을 생산하는 12개 기업에 대해
# 1년 광고비(독립변수, x)와 매출액(종속변수, y)

# (2) 결정계수 계산

summary(adsales.lm)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-5.7539 -2.0502 -0.1639  1.4285  7.4546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.28480    2.88935   1.137    0.282
x            1.59716    0.07483  21.343 1.14e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.834 on 10 degrees of freedom
Multiple R-squared:  0.9785,  Adjusted R-squared:  0.9764
F-statistic: 455.5 on 1 and 10 DF,  p-value: 1.136e-09
```

결과 해석

- Residual standard error 에 표시된
  3.834 가 잔차표준오차 값
- Multiple R-squared 에 표시된
  0.9785 값이 결정계수 R^2 의 값
- 이는 종속변수의 12개 관측값이 갖는 총변동의 97.85% 를 하나의 독립변수로를 사용한 회귀식으로 설명할 수 있다는 것을 의미하므로 이 회귀식은 상당히 유용하다는 것을 알 수 있다.
- F-statistic은 분산분석표의 F-검정 결과

회귀 계수에 대한 검정

결과 해석에 앞서 종속변수에 대한 독립변수의 유의성 검토를 위해 필요한 단계

기울기 β 에 대한 검정을 통하여 회귀분석 모형에서 추정된 회귀계수를 검정

단순선형회귀모형

- 독립변수 x가 종속변수 y에 유의한 영향을 미치는가?
  - ✓ $H_0 : \beta = 0$(유의한 영향을 미치지 않음)
  - ✓ $H_1 : \beta \neq 0$(유의한 영향을 미침)

- 검정통계량

  (귀무가설 하에서 자유도 n-2인 t-분포를 따름)

$$t = \frac{b}{SE(b)}$$

여기서, 표준오차

$$SE(b) = \sqrt{(X'X)^{-1} \cdot MSE}$$

다중회귀모형

- 단순 회귀분석 모형에서와 유사하게 t 통계량을 이용하여 수행
  - ✓ $H_0 : \beta_i = \beta_{i0}$(많은 경우, $\beta_{i0} = 0$)
  - ✓ $H_1 : \beta \neq \beta_{i0}$

- 검정통계량

  (귀무가설 하에서 자유도 n-p-1인 t-분포를 따름)

$$t = \frac{b_i - \beta_{i0}}{SE(b_i)}$$

- 편상관계수(Partial correlation coefficient)
  - 회귀계수β_i는 독립변수 x_i를 제외한 다른 독립변수의 값이 동일할 때, x의 증가분 대비 종속변수 y의 증가분을 의미함
  - 즉, 어떤 독립변수들을 함께 고려하는가에 따라 결과가 달라질 수 있음

회귀모형의 검정 실습

R CODE

유사한 제품을 생산하는 12개 기업에 대해
1년 광고비(독립변수, x)와 매출액(종속변수, y)

 # 광고비(x)의 검정

```
summary(adsales.lm)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7539 -2.0502 -0.1639  1.4285  7.4546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.28480    2.88935   1.137    0.282
x            1.59716    0.07483  21.343 1.14e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.834 on 10 degrees of freedom
Multiple R-squared:  0.9785,  Adjusted R-squared:  0.9764
F-statistic: 455.5 on 1 and 10 DF,  p-value: 1.136e-09
```

결과 해석

- $\widehat{\beta_0} = 3.28480$, $\widehat{\beta_1} = 1.59716$ 이며,
- 표준오차 SE($\widehat{\beta_0}$) = 2.88935 , SE($\widehat{\beta_1}$) = 0.07483
- 가설검정
  - ✓ $H_0 : \beta_1 = 0$
  - ✓ $H_1 : \beta_1 \neq 0$
- 검정통계량 $t = \frac{\widehat{\beta_1} - \beta_1}{SE(\widehat{\beta_1})}$

$$\rightarrow t = \frac{1.59716 - 0}{0.07483} = 21.343$$

- 이 값의 의미는 β_1과 0의 차이가 오차의 약 21배 라는 것으로 그 차이가 매우 크다는 것을 알 수 있다.
- 실제 이 값에 대한 p-value 는 1.14×10^{-9} 보다 작으므로 귀무가설 H_0 를 기각

회귀모형의 검정 실습

R CODE

유사한 제품을 생산하는 12개 기업에 대해
1년 광고비(독립변수, x)와 매출액(종속변수, y)

 # 광고비(x)의 검정

summary(adsales.lm)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5.7539 -2.0502 -0.1639  1.4285  7.4546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.28480    2.88935   1.137    0.282
x            1.59716    0.07483  21.343 1.14e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.834 on 10 degrees of freedom
Multiple R-squared:  0.9785,  Adjusted R-squared:  0.9764
F-statistic: 455.5 on 1 and 10 DF,  p-value: 1.136e-09
```

결과 해석

- $\widehat{\beta_0} = 3.28480$, $\widehat{\beta_1} = 1.59716$ 이며,
- 표준오차 SE($\widehat{\beta_0}$) = 2.88935 , SE($\widehat{\beta_1}$) = 0.07483
- 가설검정
  - ✓ $H_0 : \beta_0 = 0$
  - ✓ $H_1 : \beta_0 \neq 0$
- 검정통계량 $t = \frac{\widehat{\beta_0} - \beta_0}{SE(\widehat{\beta_0})}$

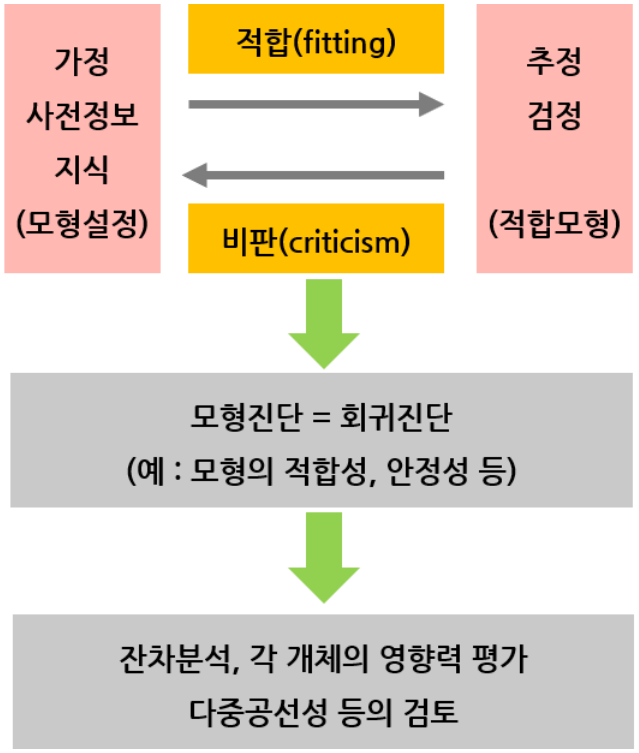
  $$\rightarrow t = \frac{3.28480 - 0}{2.88935} = 1.137$$
- p-value는 0.282 로 유의수준 $\alpha = 0.05$ 보다 크므로 귀무가설을 기각할 수 없다.
- 실제 이 값에 대한 p-value 는 1.14×10^{-9} 보다 작으므로 귀무가설 H_0 를 채택 ($\beta_0 = 0$)
- $t(0.025, 10) = 2.228$ 임을 이용하여 β_0 에 대한 95% 신뢰구간을 구해보자.
  - $3.28480 \pm (2.228) \times (2.88935) = (-3.15267, 9.72227)$
  - 신뢰구간이 0을 포함하고 있으므로 H_0 를 채택

잔차분석

회귀모형의 3가지 가정을 확인하기 위해서는 잔차분석이 필요하다.

오차항의 가정

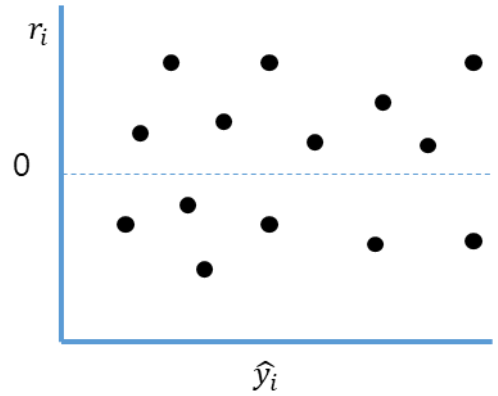
- **모형 설정**
  - ✓ 선형 모형 : $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$ 의 형태
  - ✓ $\varepsilon_i \sim N(0, \sigma^2)$을 가정
  - : 오차항에 **등분산성**, **정규성**, **독립성**을 가정
  - ✓ 실제 오차항은 관측할 수 없으므로 ε_i 와 비슷할 것으로 생각되는 잔차 $e_i = y_i - \hat{y_i}$로 가정들을 검토
- **오차항에 대한 검토**
  - ✓ 등분산성 검토 : 잔차 산점도 이용, Breusch- Pagan 검정
  - ✓ 정규성 검토 : 정규확률 그림, Jarque-Bera 검정
  - ✓ 독립성 검토 : Durbin-Watson 검정



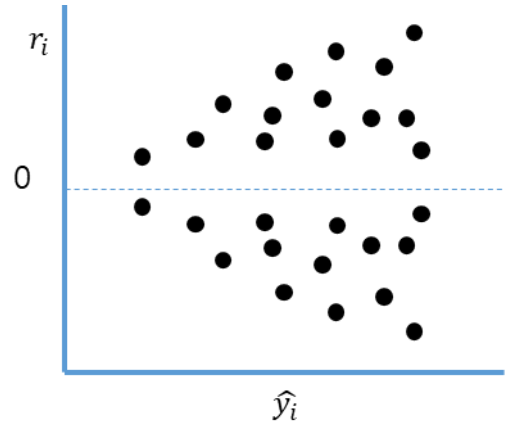
잔차분석

회귀모형의 3가지 가정을 확인하기 위해서는 잔차분석이 필요하다.

잔차산점도



- ✓ 가장 이상적인 잔차 산점도
- ✓ 잔차가 0에 대해 대칭이고 (평균 0)
- ✓ 랜덤하게 흩어져 있으며 (독립성)
- ✓ 분산이 일정 (등분산성)

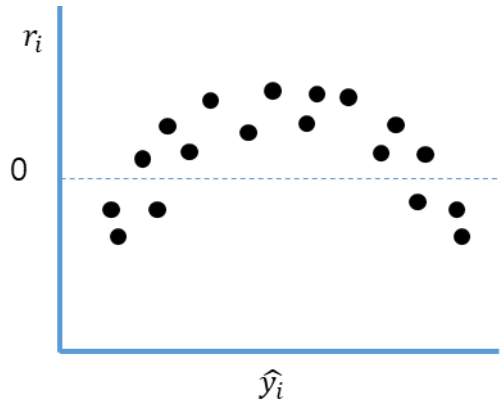


- ✓ 추정값이 증가함에 따라 잔차의 산포 증가
- ✓ 등분산 가정에 위배
- ✓ 변수 변환이나 가중 최소 제곱법 등을 이용하여 해결

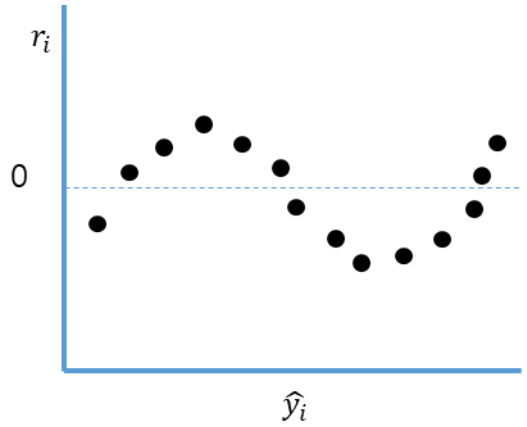
잔차분석

회귀모형의 3가지 가정을 확인하기 위해서는 잔차분석이 필요하다.

잔차산점도



- ✓ 잔차가 이차원 곡선의 형태를 보이고 있음
- ✓ 비선형회귀를 고려하거나 변환을 통해 해결

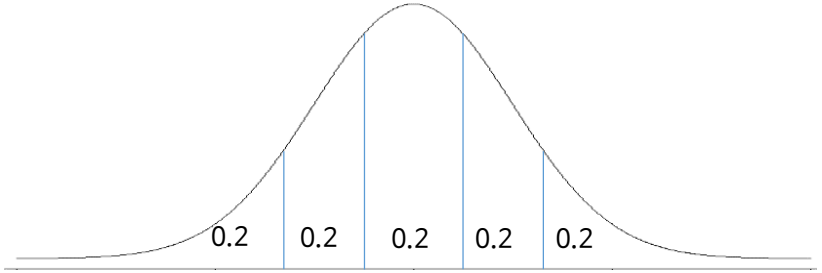


- ✓ 잔차의 형태가 0을 기준으로 양, 음의 값으로 번갈아 나타나고 있음
- ✓ 자기상관관계를 보임
- ✓ 독립성에 위배됨

잔차분석

회귀모형의 3가지 가정을 확인하기 위해서는 잔차분석이 필요하다.

정규성 : 정규확률그림(Q-Q plot)



▪ Q-Q plot 그리기

1) $e_1, \cdots, e_n$의 표본 평균이 0, 표본 분산이 1이 되도록 표준화하여 $e_1', \cdots, e_n'$ 을 얻음

2) $e_1', \cdots, e_n'$ 을 작은 순서대로 배열하고 이를 $e_{(1)}', \cdots, e_{(n)}'$ 이라고 정의

3) 만약 $e_1, \cdots, e_n$ 이 정규분포를 따랐다면 예를 들어 n이 4인 경우 $Z \sim N(0,1)$일 때,

$P(Z \leq e_{(1)}') \approx 0.2, \cdots, P(Z \leq e_{(4)}') \approx 0.8$ 일 것임

따라서 $(P(Z \leq e_{(1)}'), 0.2), \cdots, (P(Z \leq e_{(4)}'), 0.8)$ 를 좌표 평면에 y=x 직선에 가까우면 정규성을 가정할 수 있음

잔차분석

회귀모형의 3가지 가정을 확인하기 위해서는 잔차분석이 필요하다.

정규성 : 정규확률그림(Q-Q plot)

- **정규성 검정**
  - ✓ 정규확률 도표는 정규성을 평가하는 하나의 방편으로 절대적인 기준은 아님
  - ✓ 잔차의 히스토그램이나 점도표를 그려서 정규성 문제를 검토하기도 함
  - ✓ 정규성을 검정하는 방법
    - : Shapiro-Wilk test, Anderson-darling test, Kolmogorov-Smirnov test 등

- **정규성 가정을 충족하지 못할 때**
  - ✓ 정규성 가정을 충족하지 못한 경우 변수변환 등을 통해 해결하는 방안을 고려해 볼 수 있음

잔차분석

회귀모형의 3가지 가정을 확인하기 위해서는 잔차분석이 필요하다.

독립성 : Durbin-Watson 검정

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

▪ 성질

✓ D는 0~4 사이의 값

: D=4에 가까울 수록 음의 자기 상관

: D=0에 가까울 수록 양의 자기 상관

: D=2에 가까울 수록 독립성

▪ Durbin-Watson 검정 통계량

✓ 회귀분석에서 오차항의 자기상관성 여부를 대수적 방법으로 검정하기 위해 사용되는 방법

✓ 1차의 자기상관계열 가정

▪ 제한성

✓ 1차 자기상관관계열을 가정하기 때문에 오차항의 시간 종속성이 2차 이상인 경우에는 유용한 정보를 제공한다고 할 수 없음

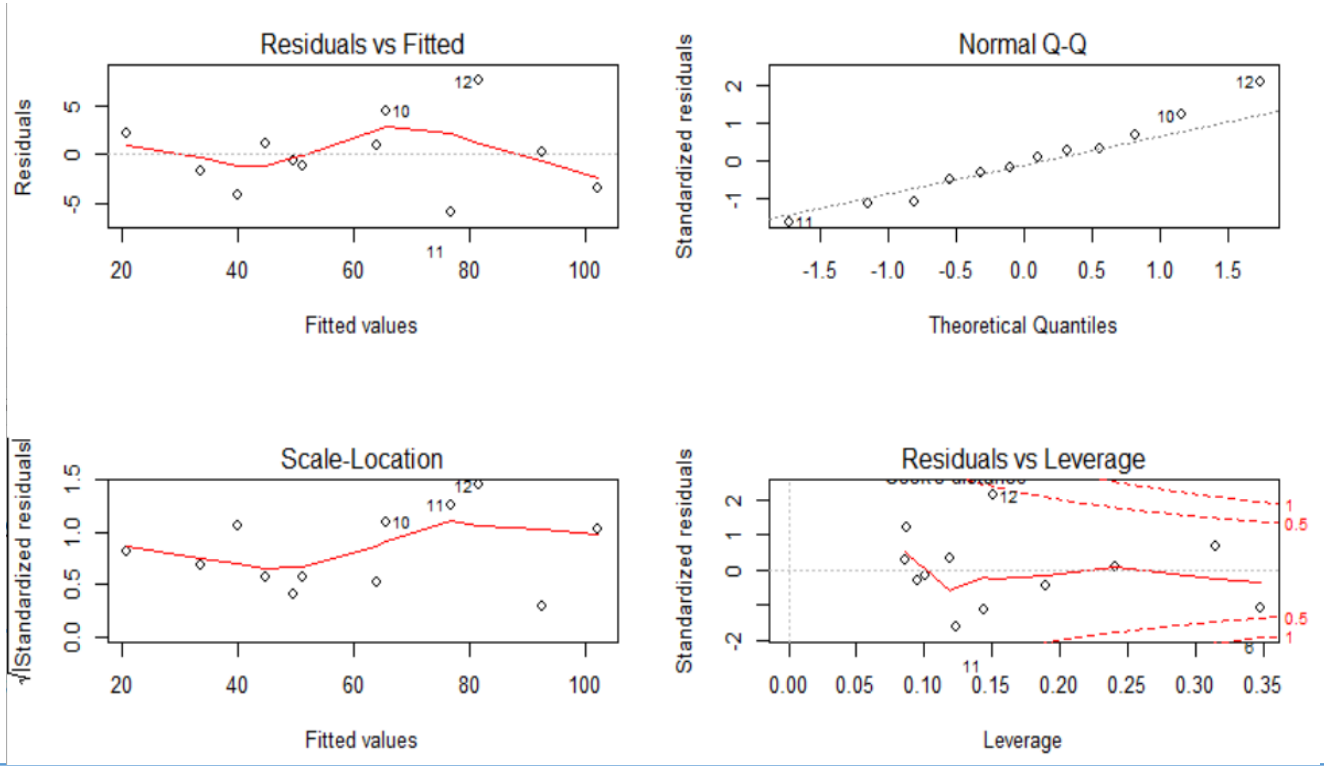
잔차분석 실습

R CODE

유사한 제품을 생산하는 12개 기업에 대해
1년 광고비(독립변수, x)와 매출액(종속변수, y)

 # 표준화 잔차들의 산점도와 정규확률도

```
par(mfrow=c(2,2))
plot(adsales.lm)
```



잔차분석 실습

결과 해석

(1) 등분산성 검정

- 상단 왼쪽의 그래프는 '잔차 대 적합값'의 산점도
- 하단 왼쪽의 그래프는 '표준화잔차의 절대값의 재곱근 대 적합값 ' 의 산점도
- 자료들에 대한 평활곡선이 함께 표시 됨

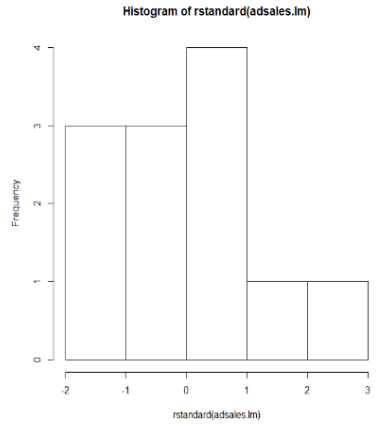
- 잔차의 절대값이 큰 2개의 자료 (10, 12 번)에 번호가 병기

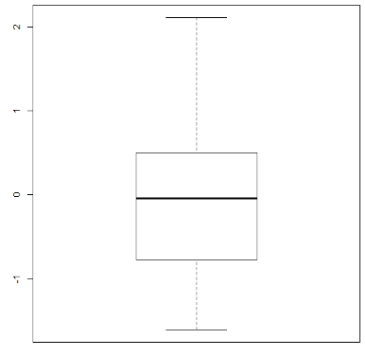
```
# 잔차
resid(adsales.lm)

# 표준화잔차
rstandard(adsales.lm)
```

결과 해석

(2) 정규성 검정

- 정규성 검정을 위해서는 상단 오른쪽의 정규확률도를 이용
- 점들이 직선에서 약간 벗어나 있어 정규성을 완전히 만족한다고 볼 수는 없으나, 이 정도의 편차는 허용되어도 무방
- 히스토그램, 상자그림 등을 이용하여 자료의 비대칭성 등을 확인





잔차분석 실습

결과 해석

(3) 독립성 검정

```
library(car)
durbinWatsonTest(adsales.lm)
```

lag	Autocorrelation	D-W Statistic	p-value
1	-0.4398787	2.470312	0.558

Alternative hypothesis: rho != 0

- DW 통계량 값이 2에 가까울 수록 오차항은 독립이다.
- 분석 결과 잔차에는 순차 상관 관계가 있다는 증거는 없다 (p-value = 0.558)

결과 해석

End of Document