다변량분석

회귀분석

❖ 다양한 선형회귀모형 종류

Syntax	선형모형	모형에 대한 설명
Y ~ A	$Y = \beta_0 + \beta_1 A + \varepsilon$	Y절편과 기울기
Y ~ -1 + A 또는 Y ~ 0 + A	$Y = \beta_1 A + \varepsilon$	원점(0,0)을 지나는 직선
Y ~ A + I(A^2)	$Y = \beta_0 + \beta_1 A + \beta_2 A^2 + \varepsilon$	2차항을 포함한 회귀
Y ~ A + B	$Y = \beta_0 + \beta_1 A + \beta_2 B + \varepsilon$	A와 B 1차항만 포함
Y ~ A:B	$Y = \beta_0 + \beta_1 AB + \varepsilon$	A와 B의 1차 상호작용만 포함
Y ~ A*B	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB + \varepsilon$	A와 B의 1차항과 상호작용 포함
Y ~ (A + B + C)^2	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 BC + \varepsilon$	A, B, C 1차항과 ()^n의 n차 상호작용까지 모두 포함한 모형

회귀분석

❖ 회귀모형 [잔차분석]

➢ 1. 회귀함수의 선형성

- 선형회귀함수가 적합한지 적합하지 않은지에 대한 문제는 자료의 산점도로 확인
- 즉, 잔차를 각 설명변수(X) 또는 반응변수(y)와의 산점도를 통해 확인

➢ 2. 오차항의 등분산성과 독립성

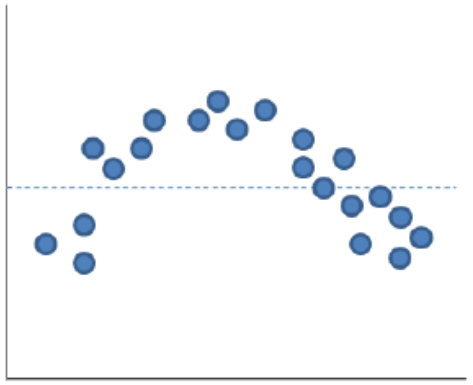
- 보통 적합된 값(fitted values)과 잔차와의 산점도를 통해 0을 중심으로 랜덤한 형태를 만족해야 함
- 잔차에 대한 독립성 검정은 Durbin-Watson 검정통계량을 통해 확인 가능

➢ 3. 오차항의 정규성

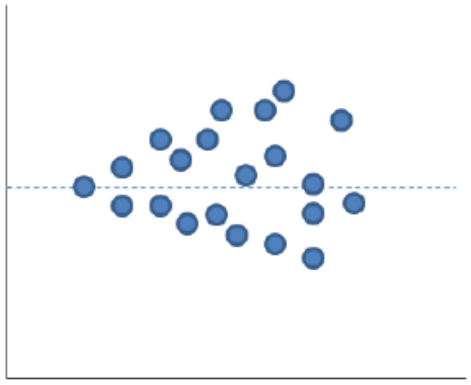
- 잔차 히스트그램 확인
- 잔차들에 대해 정규 Q-Q 그림을 통해 확인
- 정규성을 만족한다는 귀무가설에 대한 Shapiro-Wilks 검정통계량을 통해 확인

회귀분석

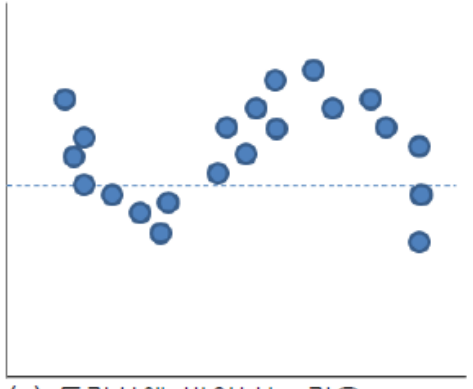
❖ 그래프를 이용한 잔차분석



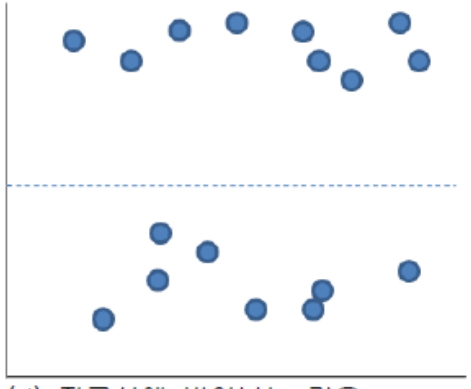
(a) 선형성이 벗어나는 경우



(b) 등분산성이 벗어난 경우



(c) 독립성에 벗어나는 경우



(d) 정규성에 벗어나는 경우

회귀분석 실습

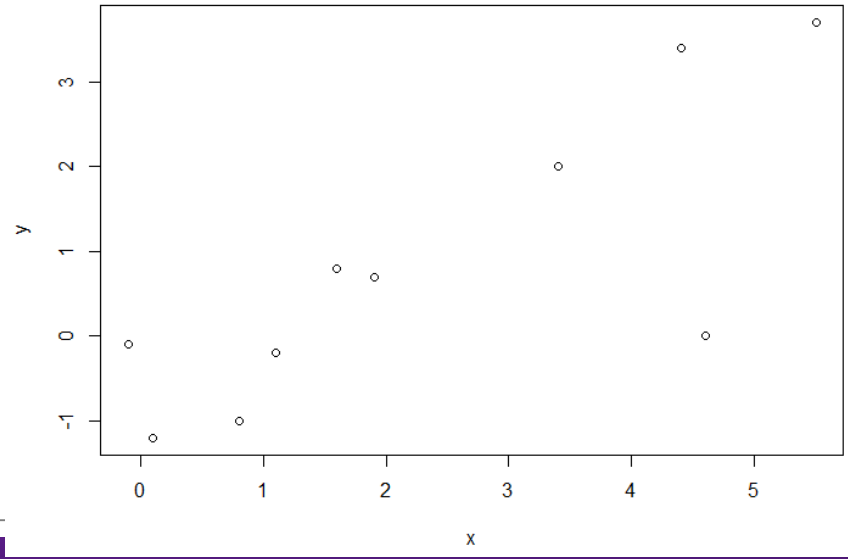
❖ 실습 1

➢ 이변량 데이터 (x, y)가 다음과 같은 경우

- x <- c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 4.6, 1.6, 5.5, 3.4)
- y <- c(0.7, -1.0, -0.2, -1.2, -0.1, 3.4, 0.0, 0.8, 3.7, 2.0)

➢ 분석 전, 변수들간의 선형관계 파악하기

- 상관관계 분석 또는 산점도 살펴보기 (선형관계 파악하기)
- plot(x,y)
- cor(x,y)



회귀분석 실습

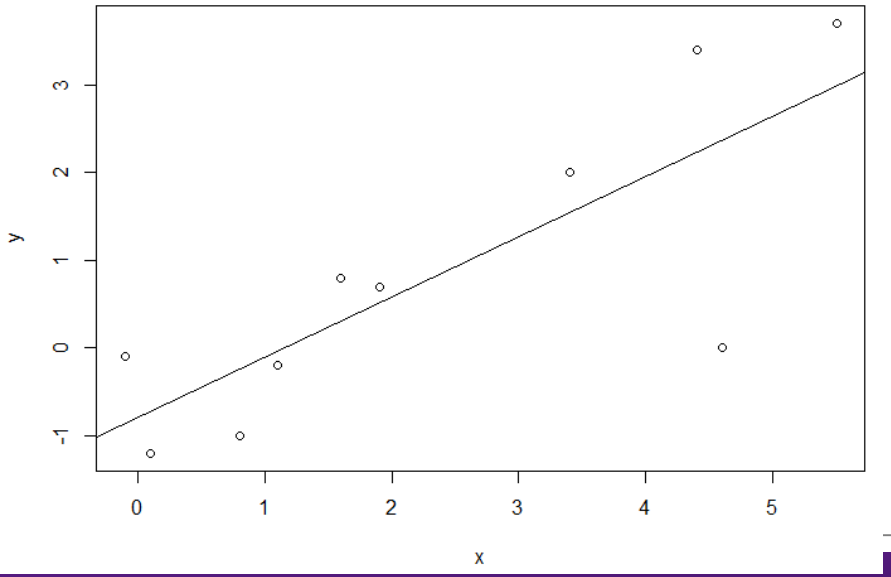
❖ 실습 1

- 회귀분석 진행
  - out <- lm(y ~ x)
  - summary(out)
- 그림으로 표현
  - plot(x,y)
  - abline(out)

```
> summary(out)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3651 -0.4036  0.3208  0.6613  1.1720

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7861     0.5418  -1.451  0.18485
x             0.6850     0.1802   3.801  0.00523 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.083 on 8 degrees of freedom
Multiple R-squared:  0.6436,    Adjusted R-squared:  0.599
F-statistic: 14.45 on 1 and 8 DF,  p-value: 0.005231
```

회귀계수의 유의성 검정

회귀모형에 대한 적합성 검정



회귀분석 실습

❖ 실습 1

➢ 추정된 회귀식 표현하기

- ?

➢ 적합된 데이터 및 예측하기

- pred_y <- predict(out, newdata = data.frame(x=x))
- out$fitted.values #값이 동일한지 확인하기

- 새로운 데이터로 예측
- pred_1 <- predict(out, newdata= data.frame(x=2.3))
- pred_2 <- predict(out, newdata= data.frame(x= c(1, 2.2, 6.7)))

회귀분석 실습

❖ 실습 2

➢ 책의 부피(독립, volume)와 무게(종속, weight)에 대해 회귀분석을 실시하시오.

- volume <- c(412, 953, 929, 1492, 419, 1010, 595, 1034)
- weight <- c(250, 700, 650, 975, 350, 950, 425, 725)
- plot(volume, weight)
- 회귀분석 실시 및 결과 (lm함수와 summary 함수이용)

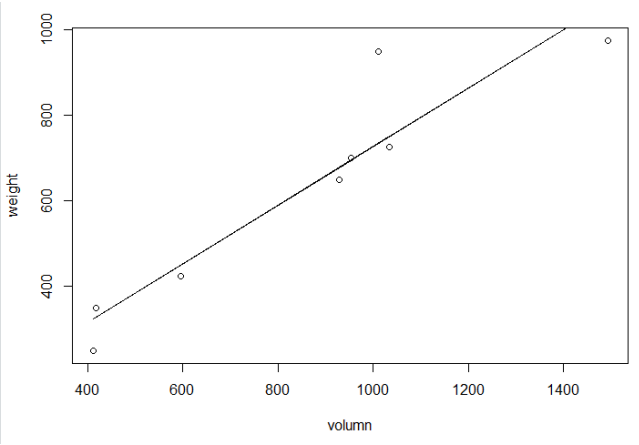
```
> book.lm <- lm(weight ~ volumn)
> summary(book.lm)

Call:
lm(formula = weight ~ volumn)

Residuals:
    Min      1Q  Median      3Q     Max
-89.674 -39.888 -25.005   9.066 215.910

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.3725    97.5588   0.424 0.686293
volumn        0.6859     0.1059   6.475 0.000644 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102.2 on 6 degrees of freedom
Multiple R-squared:  0.8748,    Adjusted R-squared:  0.8539
F-statistic: 41.92 on 1 and 6 DF,  p-value: 0.0006445
```

```
> anova(book.lm)
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
volumn     1 437878  437878  41.923 0.0006445 ***
Residuals  6  62669   10445
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

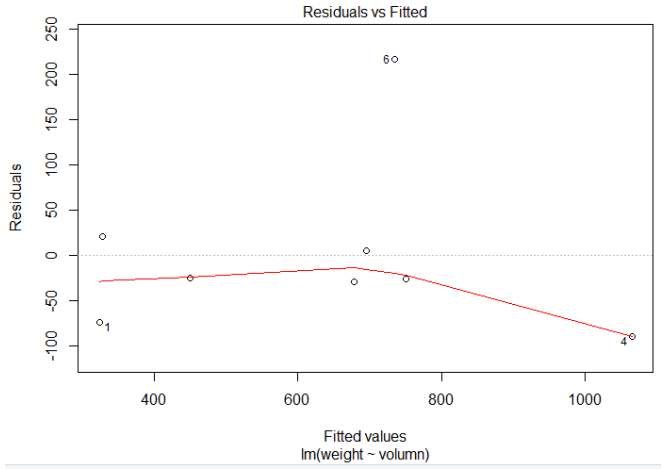
회귀분석 실습

❖ 실습 2

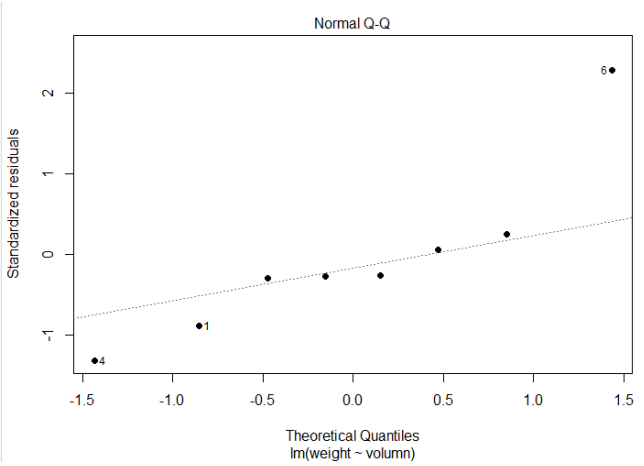
- 회귀식 그림으로 표현
  - plot(volume, weight)
  - lines(volume, book.lm$fitted.values)



- 잔차분석 그래프



plot(book.lm, which=1)



plot(book.lm, which=2, pch=16)

회귀분석 실습

D.I.Y 1

- R에 내장된 데이터 women 이용하여 분석하시오.
  - data(women)
  - women
  - weight 종속, height 독립변수

```
> data(women)
> women
   height weight
1      58    115
2      59    117
3      60    120
4      61    123
5      62    126
6      63    129
7      64    132
8      65    135
9      66    139
10     67    142
11     68    146
12     69    150
13     70    154
14     71    159
15     72    164
```

- 물음에 답하시오.
  - weight에 대한 height의 산점도를 그리시오.
  - 단순선형회귀직선을 구하여 산점도에 함께 나타내시오.
  - 모형 적합성에 대해 검정하시오,
  - 회귀계수가 유의한지 검정하시오.
  - 결정계수를 구하고 해석하시오.
  - 잔차에 대해 독립성, 등분산성, 정규성을 만족하는 잔차그림과 Q-Q 그림을 그리고 설명하시오.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , i = 1, 2, \cdots, n$$

회귀분석 실습

❖ D.I.Y 2 (다중회귀분석)

➢ 내장된 데이터 airquality 를 활용하여 회귀분석을 실시하시오.

- data(airquality)
- airquality [,1:4] 데이터를 사용
- Ozone 종속변수이며, Solar.R (일조량), Wind(풍속), Temp(기온) 은 독립변수

➢ 분석하기

- 각 변수들간의 상관분석 및 산점도 그리기
- 3개의 독립변수를 활용하여 다중회귀분석을 실시
- 모형 적합성 및 회귀계수의 유의성 검정하기
- 잔차분석(그래프 그리기: 히스토그램, 적합값과 잔차 산점도, Q-Q plot)

회귀분석 과제

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 데이터 소개

- 보스턴 지역 주택 가격 데이터 (MASS 패키지의 Boston 데이터)

```
library(MASS)
data <- Boston
```

- 변수
  - crim: 범죄발생률
  - zn: 주거지 중 2500 ft^2 이상 크기의 대형주택이 차지하는 비율
  - indus: 소매상 이외의 상업지구의 면접 비율
  - chas: 찰스강과 접한 지역은 1, 아니면 0인 더미변수
  - nox: 산화질소 오염도
  - rm: 주거지당 평균 방 개수
  - age: 소유자 주거지 중 1940년 이전에 지어진 집들의 비율
  - dis: 보스턴의 5대 고용중심으로부터의 가중 평균 거리
  - rad: 도시 순환 고속도로에의 접근 용이 지수
  - tax: 만달러당 주택 재산세율
  - ptratio: 학생-선생 비율
  - black: 흑인(BK) 인구 비율이 0.63과 다른 정도의 제곱 (1000(BK − 0.63)^2
  - lstat : 저소득 주민들의 비율 퍼센트
  - medv: 소유자 주거지 주택 가격

회귀분석 과제

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 데이터 나누기

- 학습 데이터로 모형 적합
- 검증 데이터로 모형을 평가, 비교하고, 최종 모형 선택
- 테스트 데이터로 최종모형의 일반화 능력 평가

➢ 분석 절차

- 데이터 타입, 기초 통계량, 시각화 등 탐색적 데이터 분석 실시 (일변량 데이터)
- 수치형 변수들간의 상관분석 진행 (이변량 데이터)
  - 종속변수와 독립변수들 간의 관계를 파악
  - 독립변수들 간의 관계 파악
- 모형 평가 (학습과 검증 데이터로 판단)
  - 회귀 모형의 적합성, 회귀계수의 유의성 검정, 잔차평가, 정확도(RMSE) 평가 등
- 다양한 회귀모형와 비교분석
  - 이차항을 고려한 선형회귀, 나무 모형, 랜덤 포레스트, 부스팅 등

회귀분석 과제

❖ 부동산 가격 예측 (보스턴 지역 주택 가격 데이터 활용)

➢ 데이터 나누기

- set.seed(1606)
- n <- nrow(data)
- idx <- 1:n
- training_idx <- sample(idx, n * .60)
- idx <- setdiff(idx, training_idx)
- validate_idx <- sample(idx, n * .20)
- test_idx <- setdiff(idx, validate_idx)

- training <- data[training_idx,]
- validation <- data[validate_idx,]
- test <- data[test_idx,]