R Basic Summary

Table of Contents

데이터 구조

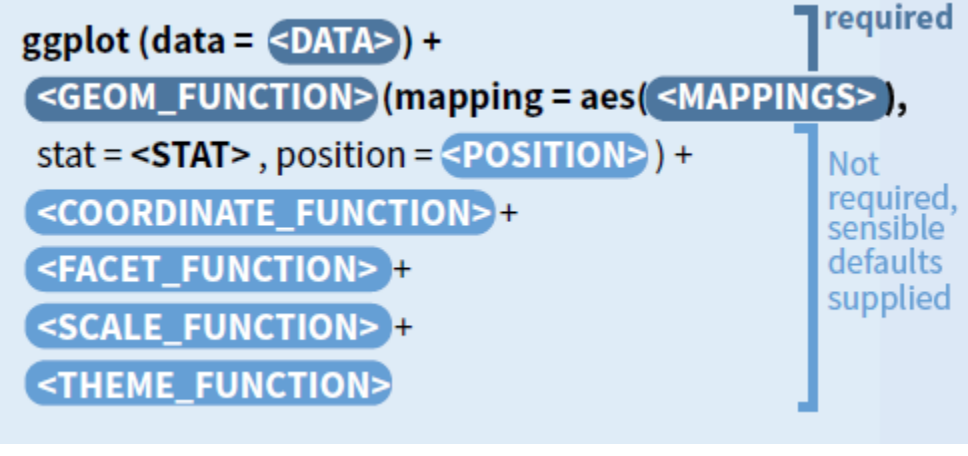
- R의 데이터 구조
  - 정수,실수,문자,문자열등 값 하나를 스칼라(scala) 라고 함
  - vector : 한가지 변수타입으로 만 구성되며 R에서의 데이터 기본단위
  - matrix: 2차원 벡터
  - data.frame : 벡터의 조합
  - array : 2차원 이상의 벡터
  - list : 다차원으로 서로 다른 데이터 구조 포함 {vector, array, data.frame, list }

데이터 가공하기 [데이터 전처리]

- 분석에 적합하게 데이터를 가공하는 작업을 데이터 전처리라고 함
- dplyr 패키지 : install.packages("dplyr")
  - 데이터 전처리 작업에 가장 많이 사용하는 패키지
- 주요 함수
  - filter(data,conditon1,con2) / filter(data,con1 | con2) : 조건으로 행 추출
  - select(data, var1, var2) : 열(변수) 추출
  - arrange(col1, col2) : 정렬
  - mutate(data, new_col=function) : 새로운 열 추가(기존+새로운 변수)
  - summarise(data, funtion) : 통계치 산출
  - group_by(factor_col) : 집단별로 나누기
  - left_join() : 데이터 합치기(열) / bind_rows() : 데이터 합치기(행)

ggplot2 패키지

- Hadley Wickham이 개발한 시각화 패키지로, R에서 가장 많이 쓰이는 패키지
- http://crantastic.org
- Grammer of Graphics(그래픽의 문법) 방법론을 R의 그리드 그래프 시스템에 적용하여 만든 패키지
- 기본 템플릿



ggplot2 패키지

- ggplot()과 <GEOM_FUNCTION> 부분만 필수적으로 기재하면 되고, 나머지는 선택사항이다
- 각각의 항목들은 괄호 처리되며, 다른 항목들과는 +로 연결된다.
- ggplot2의 레이어 구조
  - 1단계: 배경설정(축) – 필수 메인함수:ggplot( )
  - 2단계: 그래프 추가(점,막대,선) – 그래프그리기함수
  - 3단계: 설정추가(축 범위,색 표시)

산점도 / 산포도 geom_point( )
선 그래프 geom_line( )
박스플롯 geom_boxplot( )
히스토그램 geom_histogram( )
막대 그래프 geom_bar( )

- ggplot() 과 qplot() 비교
  - qplot()은 기능은 많지만 문법이 간단해서 주로 전처리 단계에서 데이터를 빠르게 활용하는 용도로 많이 쓰이고 최종적으로 결과를 보고하기 위해서는 ggplot()을 많이 사용함

텍스트 마이닝

- 텍슽마이닝이란?
  - 문자로 된 데이터에서 가치 있는 정보를 얻어내는 것
  - 가장 먼저하는 것은 문장이 어떤 품사로 되어 있는지 파악하는 형태소 분석이다.
  - 명사,동사,형용사등의 의미를 지닌 품사들의 단어들을 추출하여 각 단어의 빈도수를 확인한다.
- 워드 클라우드 만들기
  - 워드 클라우드는 단어의 빈도를 구름 모양으로 표현한 그래프이다.
  - 단어의 빈도에 따라 글자의 색과 크기가 달라짐

Table of Contents

확률함수

- 특정 분포로부터 난수를 발생시켜서 이를 통해 확률포본(random sample)을 생성
  - simulation 에 활용
  - 종류(d/p/q/r) + 확률함수
    - 확률함수의 종류
      - d: 확률밀도함수(density)
      - p: 누적확률(probability)
      - q: 4분위수(quantile)
      - r: 난수발생(random number)

```
> qnorm(0.05, mean=0, sd=1)
[1] -1.644854
> qnorm(0.05, mean=0, sd=1)
[1] -1.644854
> qnorm(0.95, mean=0, sd=1)
[1] 1.644854
> dnorm(0, mean=0, sd=1)
[1] 0.3989423
```

확률함수

확률 함수	
Function	**Description**
dnorm(x)	정규밀도함수 (default m=0 sd=1)
pnorm(q)	누적 정규 확률 (area under the normal curve to the right of q)
qnorm(p)	normal quantile 즉, 정규분포 상의 p percentile의 값
rnorm(n, m=0,sd=1)	n 개의 정규편차 (random normal deviates) (평균: m, 표준편차: sd).
dbinom(x, size, prob) pbinom(q, size, prob) qbinom(p, size, prob) rbinom(n, size, prob)	이항분포 (size = 표본 수, prob = 확률)

dpois(x, lamda)
ppois(q, lamda)
qpois(p, lamda)
rpois(n, lamda)
	poisson 분포 (m=std=lamda)
	# lamda=4일 때의 0,1, or 2 event가 발생할 확률
	dpois(0:2, 4)

dunif(x, min=0, max=1) punif(q, min=0, max=1) qunif(p, min=0, max=1) runif(n, min=0, max=1)	일양분포 (uniform distribution) #10 uniform random variates x <- runif(10)

통계분석 기법

- 통계분석
  - 기술통계 : 데이터 요약 설명.(예)전체월급 평균 구하기
  - 추론통계 : 어떤 것이 발생할 확률을 계산하는 것
- 통계적 가설검정
  - 유의확률을 이용하여 가설을 검정하는 방법
  - 유의확률:실제로는 집단간 차이가 없지만 우연히 차이가 있는 데이터라 추출될 확률. 이 값이 크다면 집단간 통계적으로 유의하지 않다.(의미없다)
  - T검정 : 두 집단의 평균에 차이가 있는지 검정.
  - 상관분석:두 변수가 관련이 있는지 검정

가설검정방법

- 통계적 가설검정
  - 표본에서 얻은 사실을 근거로 하여 모집단에 대한 가설이 맞는지 통계적으로 검정하는 분석 방법
  - 귀무가설(H0) : 직접 검정의 대상이 되는 가설 =>기각이 목표
  - 대립가설(H1) : 귀무가설이 기각될때 받아들여지는 가설 => 채택이 목표
  - 예) H0: 평균=170, H1:평균>170
  - 예) 어떤 회사의 계약직의 작년 평균 월급이 170만원, 표준편차 30만원이었다. 올해는 그것보다 높을 것이라고 생각하여 임의로 100명을 골라 평균 월급을 조사하였더니 175만원이었다. 월급이 170만원이상이라고 할 수 있는지 유의수준 a = 0.05 에서 검정하라.

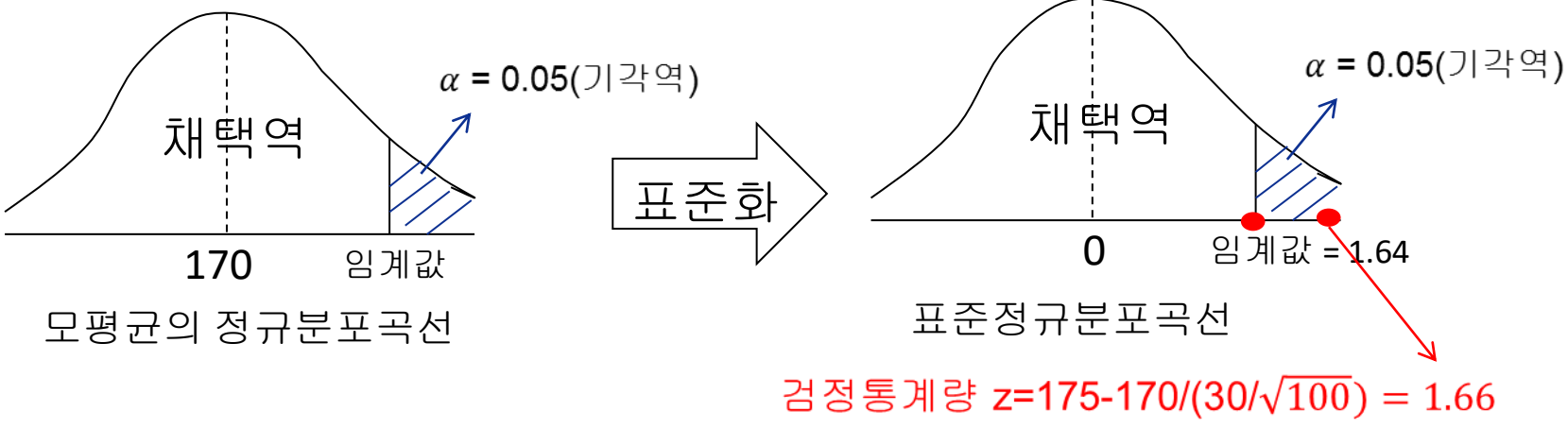
모평균=170, 모분산=30

통계적 가설

표본에서 얻은 사실, xbar=175(검정통계량)

가설검정방법

- 유의수준과 임계값(기각역의 경계값)
  - 임계값: 주어진 유의수준에서 귀무가설의 채택과 기각에 관련된 의사를 결정할때 그 기준이 되는 점
  - 유의수준(위험부담)은 귀무가설이 옳음에도 기각할 오류(제 1종오류) 임
  - 앞장예제에서 $Xbar$ = 175, μ = 170 인 모집단인 모집단에 소속되는가를 판단하는 기준점이 임계점임. H0: μ = 170 (기각 or 채택)



가설 검정

- 가설검정 = 가설 설정후 관측치를 가지고 검증
- 1종 오류 = 올바른 H를 기각하는 것
- 유의수준 significance level = type I error 의 확률 (α)
- 정규분포를 이용한 검정(norm)
  - 모평균의 Lower tail 검정(분산을 알때)
  - 모평균의 Upper tail 검정(분산을 알때)
  - 모평균의 양측검정(분산을 알때)
- T분포를 이용한 검정(t-test)
  - 모평균의 Lower tail 검정(분산을 모를때)
  - 모평균의 Upper tail 검정(분산을 모를때)
  - 모평균의 양측검정(분산을 모를때)
- 정규분포를 이용한 검정(prop.test())
  - 모집단 비율의 Lower 검정
  - 모집단 비율의 Upper 검정
  - 모집단 비율의 양측검정

2개의 모집단에 대한 추정

- 표본을 통해서 2개 모집단을 비교하는 것
- 2개 대응표본간의 모평균
  - t.test(immer$Y1, immer$Y2, paired=TRUE)
- 2개 독립표본간의 모평균 (분산분석필요)
  - var.test(mpg.auto, mpg.manual, mpg)
  - t.test(mpg.auto, mpg.manual)
- 2개 모집단 비율의 비교
  - library(MASS)
  - head(quine)
  - prop.test(table(quine$Eth, quine$Sex), correct=FALSE)

통계분석 기법(T검정)

- t.test(data=mpg_diff,cty~class,var.equal=T)
  - p-value: 유의확률.0.05미만이면 집단간 통계적으로 유의하다. 즉, 실제로는 차이가 없는데 우연히 관찰될 확률이 5%미만이라면 우연이 아니다.
- 일반휘발유와 고급휘발유 사용 자동차간 도시연비의 통계적 유의값
  - mpg_diff2=mpg %>% select(fl,cty) %>% filter(fl %in% c("r","p"))
  - table(mpg_diff2$fl)
  - t.test(data=mpg_diff2, cty~fl, var.equal = F)
    - p-value0.05보다큰 0.2875. 결국 일반휘발유와 고급휘발유간 도시연비차이가 통계적으로 유의하지 않다.

통계분석 기법(상관분석)

- 두 연속 변수가 서로 관련 있는지 검정하는 통계분석 기법
- 도출된 상관계수가 1에 가까울수록 관련성이 크다는 의미
- 상관계수가 양수이면 정비례, 음수이면 반비례한다

- 실업자수와 개인소비지출의 상관관계
  - economics =as.data.frame(ggplot2::economics)
  - cor.test(economics$unemploy,economics$pce)
    - p-value<2.2e-16 0.05보다 작으니 통계적으로 유의하다
    - cor 0.6139997 정비례관계이다.

통계분석 기법(회귀분석)

- 단순 회귀 분석
  - 설명변수들과 반응변수들 간의 함수관계를 밝히기 위한 통계적 방법이 회귀분석

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  - 광고비와 판매액 데이터를 가지고 광고비를 독립변수, 판매액을 종속변수로 할 때 기울기와 추정치를 구하여 광고비에 따른 판매액예측

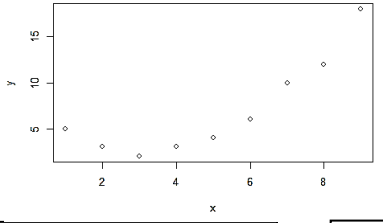
```
> x=c(4,6,6,8,8,9,9,10,12,12)
> y=c(39,42,45,47,50,50,52,55,57,60)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     28.672        2.503
```

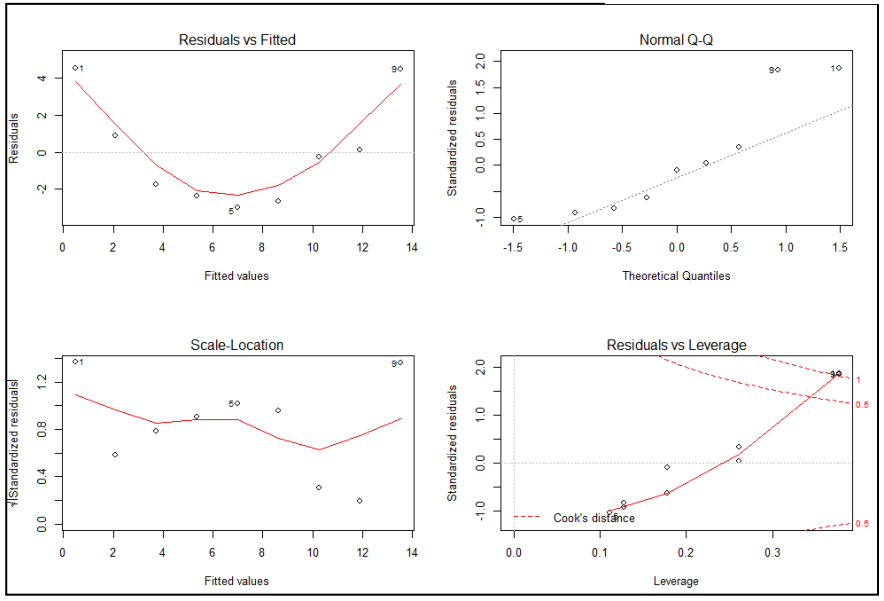
$$y = 28.67 + 2.50x$$

- 다중선형회귀분석 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$
  - lm(y ~ x1 + x2 + x3, data=z)

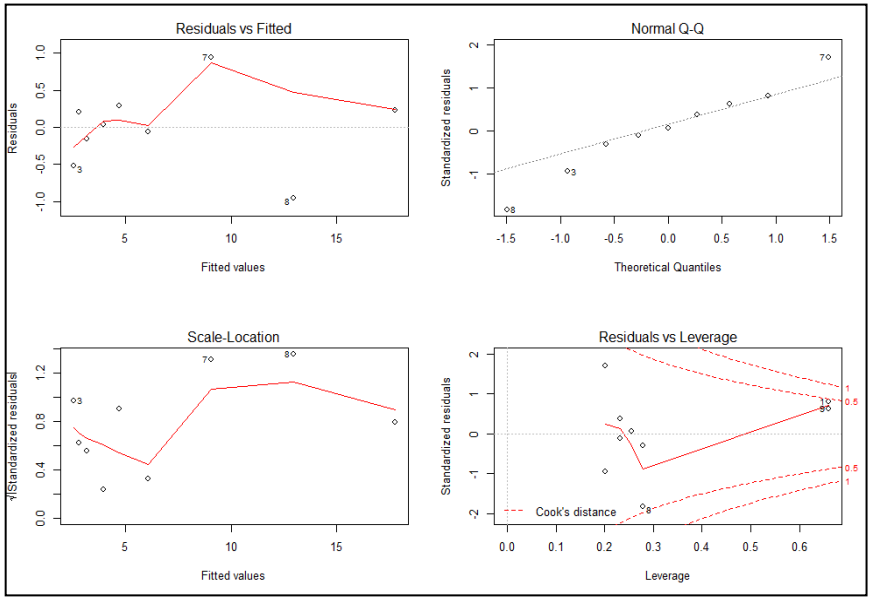
통계분석 기법(회귀분석) – 잔차분석

- lm(y~x,data=z)





잘못된 분석



잘된 분석

통계분석 기법(회귀분석)

- 회귀분석 – 고려사항
  - 선형회귀모형에 사용되는 가정들
    - 선형성 : 입력변수와 출력변수와의 관계가 선형적
    - 등분산성 : 오차의 분산이 입력 변수와 무관하게 일정
    - 정규성 : 오차의 분포가 정규분포
    - 선형 모형을 자료에 적합하기 전에 위 3가지 가정이 만족하는지 확인
    - 위 3가지 가정이 맞지 않으면 선형회귀 모형은 좋지 않은 결과를 제공
  - 가정의 검증방법
    - 선형성 : 선형회귀모형에서 입력변수와 출력변수와 산점도를 이용, 다중 선형회귀 모형에서는 잔차와 출력변수와의 산점도 이용
    - 선형회귀모형에서 사용된 모든 가정이 만족되는 경우 -> 잔차는 아무런 정보가 없는 오차와 비슷 -> 잔차의 산점도에서 어떤 패턴이 발견되면 선형회귀 모형의 가정을 의심