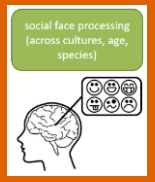
분석 인프라 활용 AI교육(R)
Day 2

Cho Heeseung

hscho9384@korea.ac.kr

목차

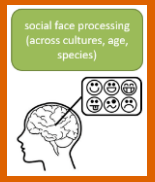


CognitiveSystems

1. 통계분석 개요
2. 기초통계 및 가설검증
3. 선형회귀분석
4. 로지스틱 회귀분석

1. 통계분석 개요

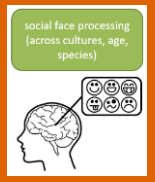
데이터 분석 프로세스



CognitiveSystems

요건 정의	모델링	검증 및 테스트	적용
분석요건 도출	데이터 마트 설계/구축	운영 테스트	운영 적용
수행방안 설계	탐색적 데이터 분석	비즈니스 영향도 평가	주기적 리모델링
	모델링		
	성능 평가		

요건 정의

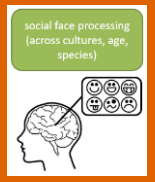


CognitiveSystems

1. **분석목표 도출**
   분석 업무의 배경을 이해하고 관계자들과 의사소통을 통해 분석 현황과 요건을 식별
   ex) 서비스의 고객 이탈 원인 분석 및 이탈 방지 방안 모색,
   공정 내 제품 불량 예측 및 불량 원인 파악
   As-is, To-be 설계

2. **수행방안 설계**
   데이터 및 운영 현황 파악 후 적용 가능한 모델/알고리즘 파악 및 평가 기준 수립
   ex) DB 접근 권한 확보 및 데이터 흐름 파악, 시스템 영향도 분석
   군집모델(K-means Cluster)?
   분류모델(Logistic Regression)?
   예측모델(Random Forest)?

모델링



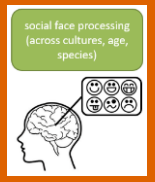
CognitiveSystems

1. **데이터마트 구축**
   데이터 수집 및 연동, 전처리 및 구조화
   ex) DB 권환 확인, ETL 등

2. **탐색적 데이터 분석(EDA)**
   분석 목표 수립 단계에서 정의한 모델링 알고리즘 적용 가능성 파악
   ex) 어떤 변수가 관련이 있을지, 새로운 파생변수 생성이 가능한지

3. **모델링 및 성능 평가**
   EDA를 통해 완성된 최종 데이터셋을 모델링 알고리즘에 적용, 평가 기준을 통해 최종 모형 결정.

검증 및 테스트

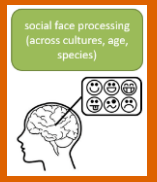


CognitiveSystems

1. **운영 테스트**
   모델링을 시스템과 유사한 환경에서 테스트, 분석 결과 및 영향도 파악
   ex)모델이 과적합 하는가? 기존 시스템에 부하를 주는가?
   1일 몇 회 이상 검증이 가능한가?

2. **비즈니스 영향도 평가**
   최종 검증 및 모형 결과 확인, 적용 후 기대 및 우려사항 정리
   ex) As-is vs To-be 결과 검토, 수익률 검토, 타 사업부문 영향 확인

적용

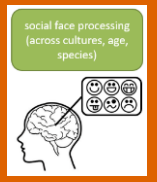


CognitiveSystems

1. **운영 적용**
   실제 시스템 내 테스트 후 배포, 실시간 혹은 스케줄러 실행
   모니터링 및 기대결과 확인

2. **주기적인 리모델링**
   한번 만든 모델이 계속 동일한 결과를 낼 수 없으므로, 주기적으로
   재평가하여 수정 및 보완

통계를 배우는 이유



CognitiveSystems

통계학: 데이터에서 의미를 찾아내는 학문

데이터 분석에 있어서 통계학은 필수 소양!

통계 기초

- **표본공간**(Sample space): 일어날 수 있는 모든 결과들의 집합
- **사건**(Event): 표본공간의 부분집합, 어떤 결과가 일어날 수 있는가?
- **확률**(Probability): 특정 사건이 일어날 가능성의 척도

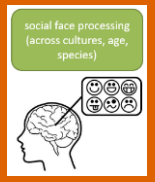
- **확률 변수**(Random Variable)
  특정 값이 나타날 가능성이 확률적으로 주어지는 변수

*** 어떤 x가 확률 변수로 주어지면, 현재의 x의 값과 다른 상황에서의 x의 값은 일반적으로 다르다.

(ex)X: 주사위를 던져서 나온 숫자, 가위바위보, ...

```
### Random Variable
sample(c(1,2,3,4,5,6),1)
sample(c(1,2,3,4,5,6),10, replace = TRUE)

mean(c(1,2,3,4,5,6))
sd(c(1,2,3,4,5,6))
```

통계 기초



CognitiveSystems

1. **이산형 확률분포**(Discrete distribution)
   : 확률 변수가 가질 수 있는 값이 셀 수 있을 경우
(ex) Uniform, Bernoulli, Poisson 등

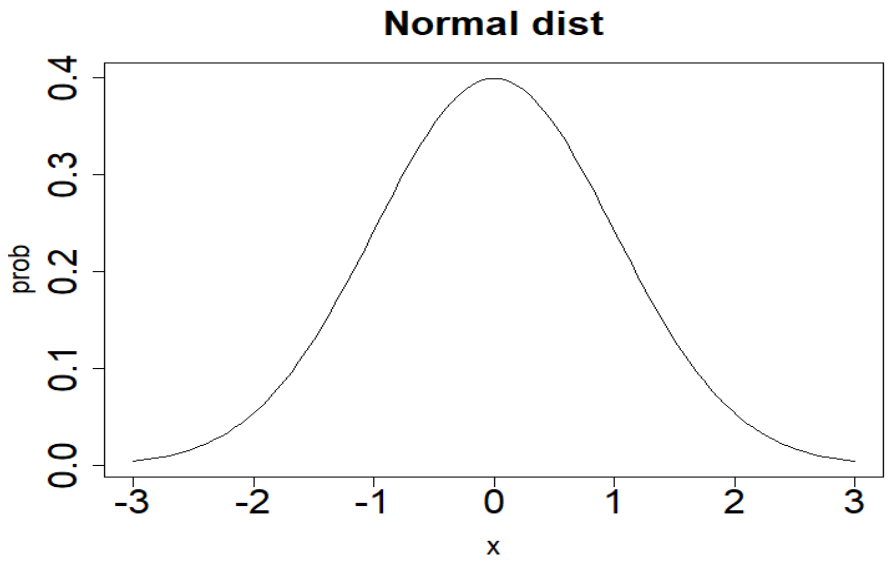
2. **연속형 확률분포**(Continuous distribution)
   : 확률 변수가 가질 수 있는 값이 셀 수 없을 경우
(ex) Uniform, Normal(Gaussian), exponential 등

정규분포

- **정규분포**는 평균이 μ, 분산이 σ^2인 연속확률분포로, 확률변수의 값이 평균에 몰려 있는 분포

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x-\mu}{2\sigma^2}\right)$$

$$X \sim N(\mu, \sigma^2)$$



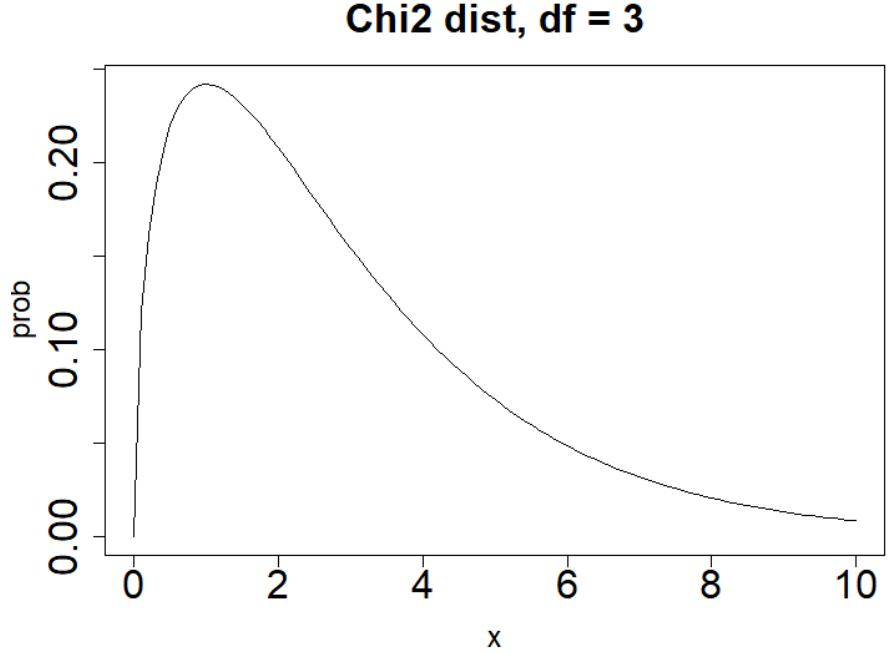
- R을 이용하여 정규분포의 값을 쉽게 찾아보자.

```
pnorm(2, mean = 2, sd = 4)          #P(X<2)까지의 확률
dnorm(0, 2, 4)                      #X=0일 때 확률밀도함수 값
qnorm(0.5, 2, 4)                    #p = 0.5가 되기 위한 확률
plot(dnorm,-3,3, main = "Normal dist", ylab = "prob", cex.main=2, cex.axis = 2, cex.lab = 1.5)
```

정규분포의 파생분포

- χ^2**분포**: 표준정규분포를 가지는 확률변수의 제곱을 더한 확률변수
  $Q = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$, k라는 자유도라는 변수를 받는다.

```
### Chi-square distribution
pchisq(2, df = 3)              #P(X<2) 까지의 확률
dchisq(2, df = 3)              #x=2일 때 확률밀도함수 값
qchisq(0.4275933, df = 3)      #p = 0.4275933이 되기 위한 확률변수값
plot_chi = function(x){
  dchisq(x, df = 3)
}
plot(plot_chi,0,10, main = "Chi2 dist, df = 3", ylab = "prob", cex.main=2, cex.axis = 2, cex.lab = 1.5)
```

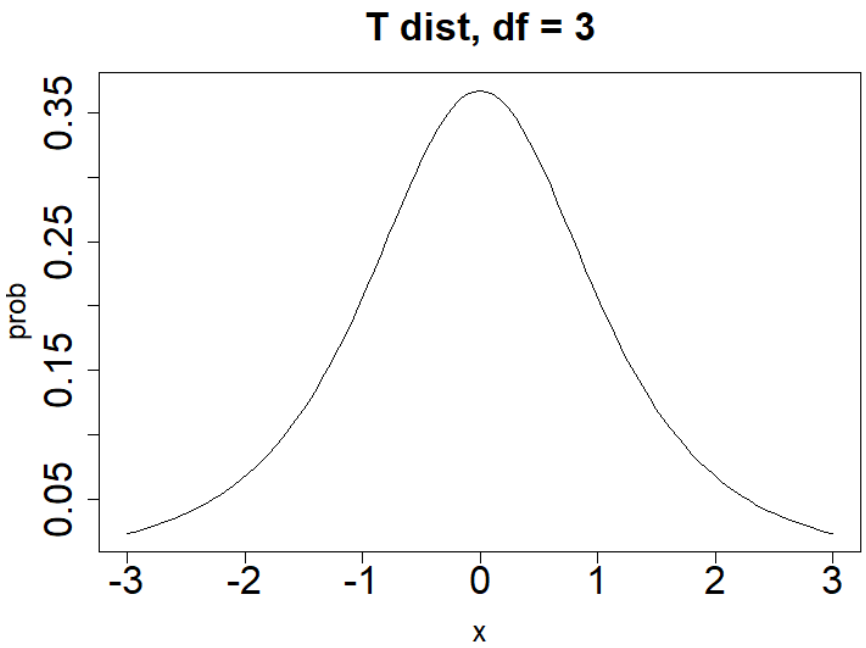


정규분포의 파생분포

- **T 분포**: 표준정규분포와 χ^2분포를 결합하여 만든 확률분포

  $T = \frac{Z}{\sqrt{Q/k}} \sim t(k),$ k라는 자유도라는 변수를 받는다.

```
### t distribution
pt(0, df = 3)                    #P(X<0) 까지의 확률
dt(0, df = 3)                    #x=0일 때 확률밀도함수 값
qt(0.5, df = 3)                  #p = 0.5가 되기 위한 확률변수값
plot_t = function(x){
  dt(x, df = 3)
}
plot(plot_t,-3,3, main = "T dist, df = 3", ylab = "prob", cex.main=2, cex.axis = 2, cex.lab = 1.5)
```

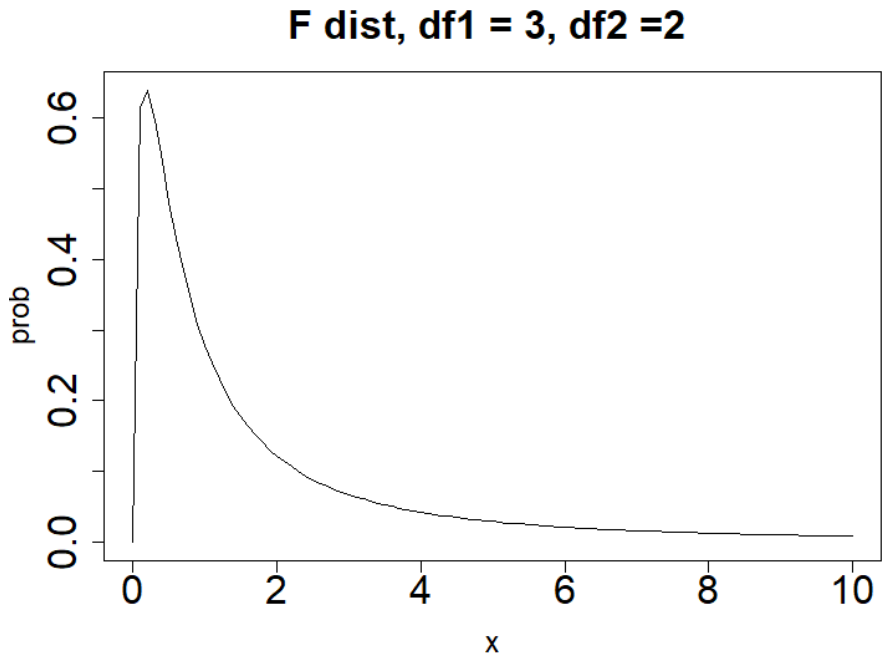


정규분포의 파생분포

- **F 분포**: 서로 다른 두개의 χ^2 분포를 결합하여 만든 확률분포

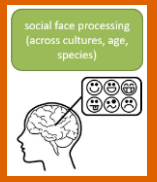
  $F = \frac{Q_1/k_1}{Q_2/k_2} \sim F(k_1, k_2)$, k1, k2라는 두개의 자유도를 변수로 갖는다.

```
### F distribution
pf(2, df1 = 3, df2 = 2)                    #P(X<2) 까지의 확률
df(2, df1 = 3, df2 = 2)                    #x=2일 때 확률밀도함수 값
qf(0.6495191, df1 = 3, df2 = 2)            #p = 0.6495191가 되기 위한 확률변수값
plot_F = function(x){
  df(x, df1 = 3, df2= 2)
}
plot(plot_F,0,10, main = "F dist, df1 = 3, df2 =2 ", ylab = "prob", cex.main=2, cex.axis = 2, cex.lab = 1.5)
```



2. 기초통계 및 가설검증

통계분석이란?



CognitiveSystems

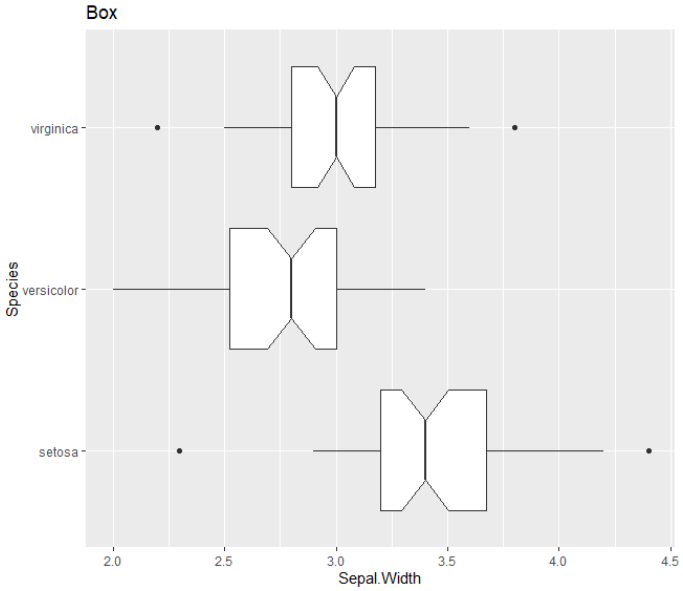
특정한 집단이나 **불확실한 현상**을 대상으로 자료를 수집해 대상 집단에 대한 정보를 구하고, 적절한 통계 방법을 이용해 의사결정을 하는 과정.

기술통계

데이터에 분석에 앞서 데이터의 대략적인 통계적 수치를 계산해 봄으로써 데이터에 대한 대략적인 이해와 앞으로 진행될 분석에 대한 통찰을 얻기 위한 분석 방법

ex) 평균, 표준편차, 그래프 확인, EDA 결과 등

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

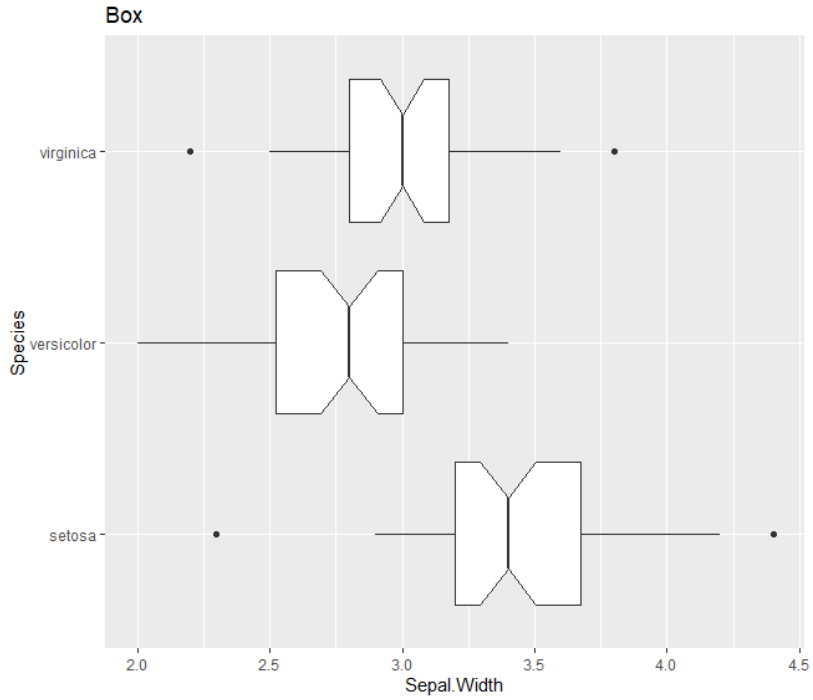


iris의 종에 따른 Sepal.Width 평균 및 분포를 시각화하여 확인

통계적 추정

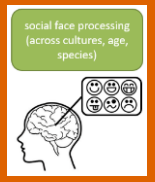
데이터로부터 얻은 내용들을 바탕으로 데이터에 대한 가설을 설정한 후, 표본 관찰을 통해 추정한 결과를 이용하여 해당 가설의 채택여부를 분석

ex) 서비스를 이탈하는 사람은 그렇지 않은 사람보다 서비스 이용시간이 적은가?



iris 종에 따라 Sepal.Width의 평균 길이가 다른 것 같다.(가설)

통계적 추정



CognitiveSystems

- **모집단**(Population): 정보를 얻고자 하는 관심 대상의 전체 집합
- **모수**(Parameter): 모집단의 특정한 성질을 나타내는 변수(평균, 분산 등)

- **표본**(sample): 모집단의 부분집합으로, 모집단으로부터 추출한 데이터
- **통계량**(estimator): 표본의 특징을 수치화한 값(표본 평균, 표본 분산 등)

- **추정**(estimation): 표본 통계량을 바탕으로 모수의 값을 예측하는 것
- **신뢰도**(Confidence level): 0~1(0~100%) 사이의 측정의 일관성을 표현

ex) **서비스 이탈하는 사람**(모집단)들의 **나이의 평균**(모수)를 추정하기 위해 일부 **사용자들의 데이터**(표본) 내의 **나이의 평균**(통계량)를 구해 전체 나이대를 **95%의 신뢰도** 내에서 예측.

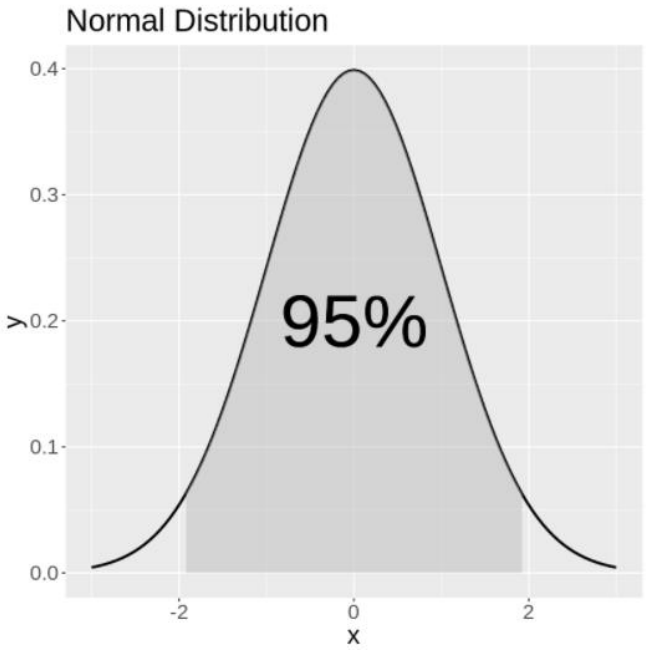
통계적 추정

구간 추정: 확률로 표현한 믿음의 정도 아래 모수가 특정한 구간에 있을 것이라 선언하는 것.

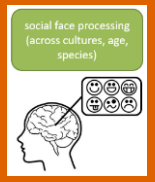
$$(\bar{X} - Z_{\alpha/2}\frac{\sigma^2}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma^2}{\sqrt{n}})$$

ex) 신뢰도 95%: Z = 1.96, 99%: Z = 2.58

```
### Confidence Interval
sample.mean = mean(iris$Sepal.Length)       #평균
sample.n = length(iris$Sepal.Length)        #표본 수
sample.sd = sqrt(sd(iris$Sepal.Length))     #분산

alpha = 0.05
degrees.freedom = sample.n - 1              #자유도
score = qnorm(p=1-alpha/2)                  #신뢰상수, 95%=1.96

margin.error = score * sample.se
lower.bound = sample.mean - margin.error
upper.bound = sample.mean + margin.error
print(c(lower.bound,upper.bound))

#Short cut
ci_Sepal.Length = lm(Sepal.Length~1, data= iris)
confint(ci_Sepal.Length, level = 0.95)    # 이때는 t분포를 따름
```



가설검증



CognitiveSystems

모집단에 대한 어떤 가설을 설정한 뒤에 표본관찰을 통해 그 가설의 채택여부를 결정하는 분석방법

- **귀무가설**(Null Hypothesis, H_0): 일반적인 통념에 해당하는 가설

$$H_0: \mu = 0 \, , H_0: \mu > 0$$

- **대립가설**(Alternative Hypothesis, H_1): 귀무가설에 반대되는 가설, 주장해야 하는 가설

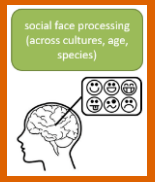
$$H_1: \mu \neq 0 \, , H_1: \mu \leq 0$$

- **검정 통계량**: 검정에 사용되는 통계량

** 귀무가설이 옳다는 전제하에 검정 통계량을 구한 후에 이 값이 나타날 가능성의 크기에 의해 귀무가설의 채택 여부를 결정

- **유의수준**(Significant level): 가능성의 크기 = 신뢰도

가설검증



CognitiveSystems

- **Type 1 error**: 귀무가설이 옳은데도 귀무 가설을 기각하게 되는 오류
- **Type 2 error**: 귀무가설이 옳지 않은데도 귀무가설을 채택하는 오류
- **유의확률**(P-value): 귀무가설이 맞다고 가정할 때 얻은 결과보다 극단적인 결과가 실제로 관측될 확률 = Type 1 error가 일어날 확률

  P-value = $P(X >$ 검정통계량$)$ or $P(X <$ 검정통계량$)$

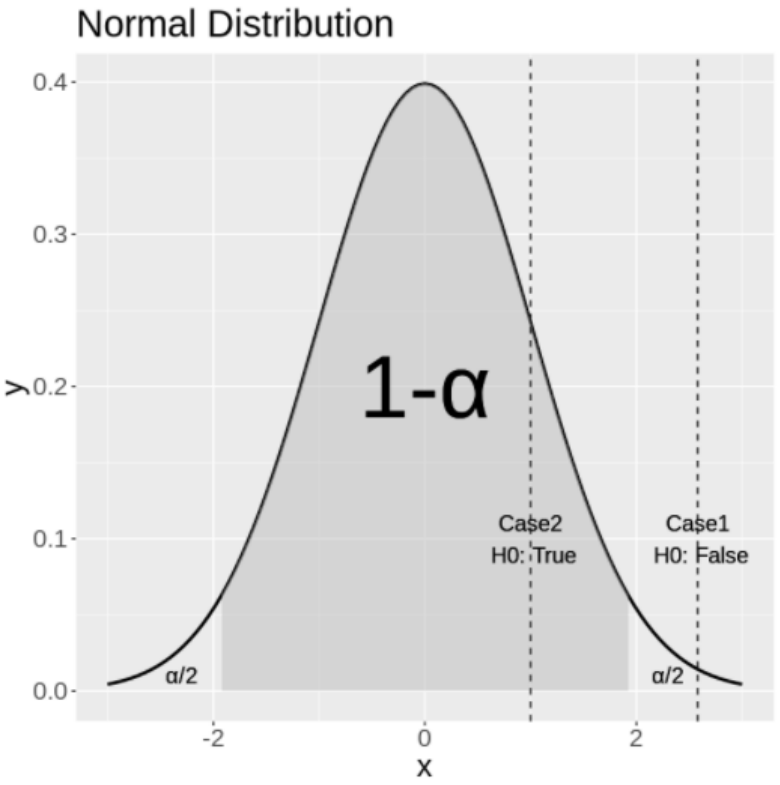
실제상황 \ 의사결정	θ_0 채택	θ_0 기각
H_0 가 사실	올바른 결정 확률=$1-\alpha$	1종오류 확률=α
H_0 가 허위	2종오류 확률=β	올바른 결정 확률=$1-\beta$=검정력

가설검증

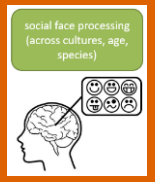
귀무가설 채택 여부

1)검정 통계량을 기반으로
Case1: 검정통계량이 Significant level 영역 외부: 귀무가설 기각
Case2: 검정통계량이 Significant level 영역 내부: 귀무가설 채택

2) p-value를 기반으로
p-value < Significant level: 귀무가설 기각
p-value > Significant level: 귀무가설 채택



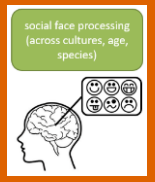
검정 방법



CognitiveSystems

1. T-test (+상관분석)
2. χ^2-test
3. F-test (+ANOVA)

T-Test



CognitiveSystems

T분포를 이용한 검정으로, 모집단의 **분포를 모를 때** 일반적으로 사용

1) 단일 표본 평균 검정 : 해당 표본평균이 얼마인가?

ex) Iris 꽃의 꽃받침 길이의 평균이 5.5인가?

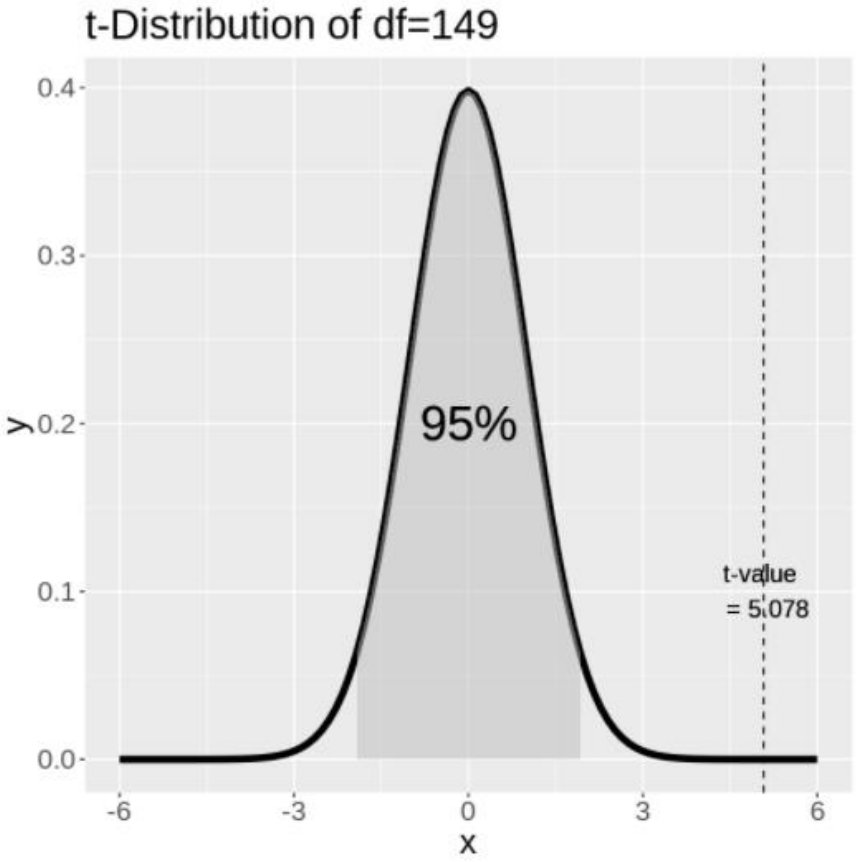
```
# equal
t.test(iris$Sepal.Length, alternative = "two.sided", mu = 5.5, conf.level = 0.95)
```

ex) Iris 꽃의 꽃받침 길이의 평균이 5.5보다 작은가?

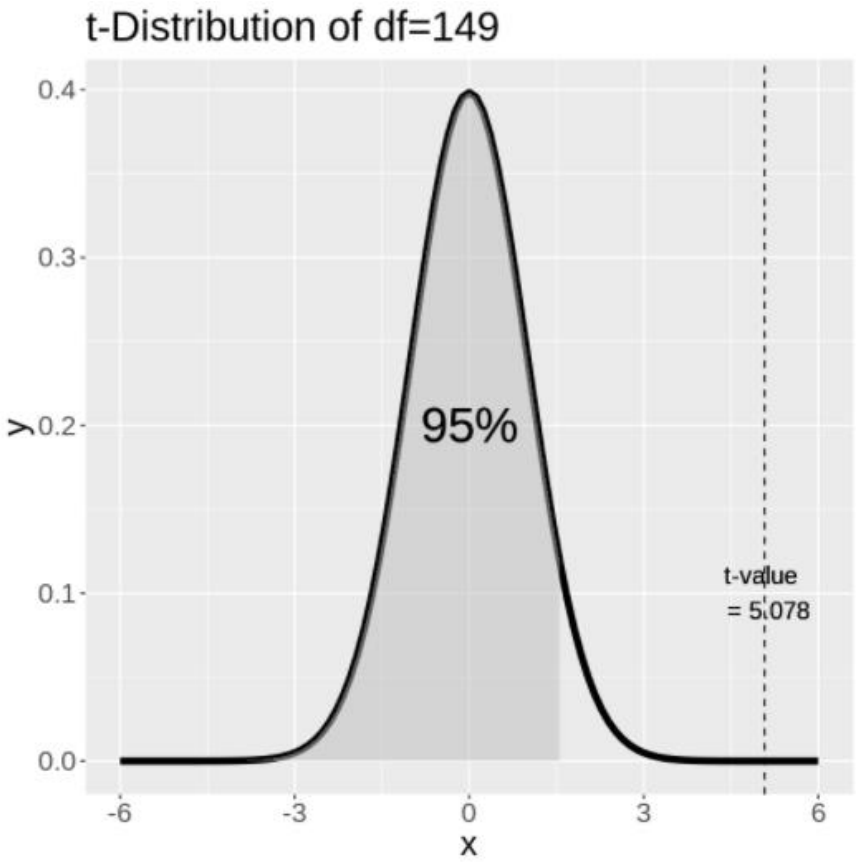
```
# greater/less than
t.test(iris$Sepal.Length, alternative = "greater", mu = 5.5, conf.level = 0.95)
```

T-Test

결과를 그림으로 알아보자.

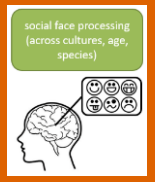


Case1



Case2

T-Test



CognitiveSystems

2) 2 표본 평균 검정: 두 표본의 평균이 같은가?

ex) Iris 꽃의 'setosa'종과 'virginica' 종의 꽃받침 길이의 평균이 같은가?

$$H_0: \mu_1 = \mu_2 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_2$$

*표본의 수가 같을 때

```
### paired t-test
iris_2group = iris[iris$Species != 'versicolor',]
t.test(Sepal.Length~Species, data = iris_2group, paired = T)
```

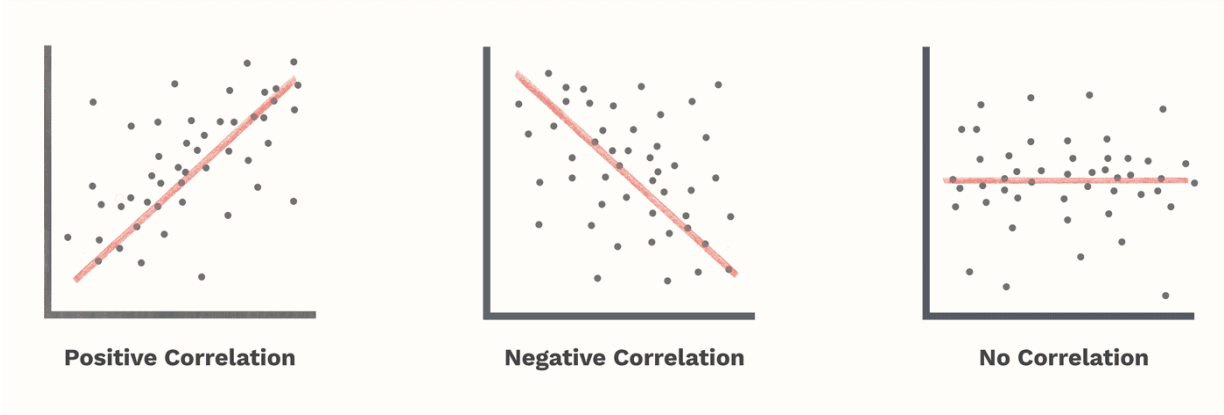
*표본의 수가 다를 때

```
### unpaired t-test
iris_setosa = iris
iris_setosa[iris_setosa$Species == 'versicolor',]$Species = 'virginica'
t.test(Sepal.Length~Species, data = iris_2group, paired = F)
```

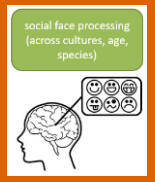
상관분석

상관계수(ρ): 데이터 내 두 변수 간의 관계를 파악하는 척도, 0~1 사이의 값
두 변수 간의 연관된 정도만 제시, 추후 두 변수간 원인과 결과의 인과관계의 방향, 정도, 모형 적합을 통한 함수 관계를 검토 가능.

(-1, -0.7)	(-0.7, -0.4)	(-0.4, -0.1)	(-0.1, 0.1)	(0.1, 0.3)	(0.3, 0.7)	(0.7, 1)
강한 음의 선형관계	뚜렷한 음의 선형관계	약한 음의 선형관계	무시될 수 있음	약한 양의 선형관계	뚜렷한 양의 선형관계	강한 양의 선형관계



상관분석



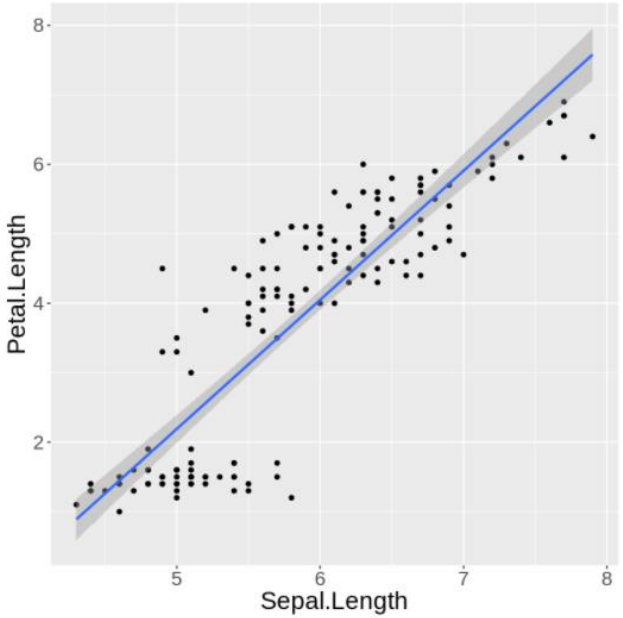
CognitiveSystems

상관계수 검정: 두 변수간의 상관관계가 있는가?

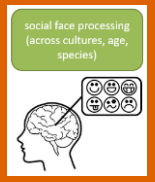
ex) 꽃받침 길이와 꽃잎 길이는 서로 상관이 있는가?

$$H_0: \rho = 0 \quad \text{vs} \quad H_1: \rho \neq 0$$

```
### Corr test
#Check Correlation
cor(iris$Sepal.Length,iris$Petal.Length)          #Correlation
cor(iris[,1:4])                                   #Corr Matrix
ggplot(iris, aes(Sepal.Length, Petal.Length))+    #Plotting
  geom_point()+geom_smooth(method = "lm")+
  theme(text = element_text(size=20))

#Cor.test
cor.test(iris$Sepal.Length,iris$Petal.Length)
```



χ^2-test



χ^2 분포를 이용한 검정으로 빈도를 이용한 값을 추정.

*주로 모집단 내 두 변수가 독립인지 검증할 때 이용.

H_0: X_1와 X_2가 독립 vs H_1: X_1와 X_2가 독립이 아니다.

```
### Chi-square Test
corr_tab = xtabs(~Sepal.Length+Petal.Length, data=iris)
chisq.test(corr_tab)
```

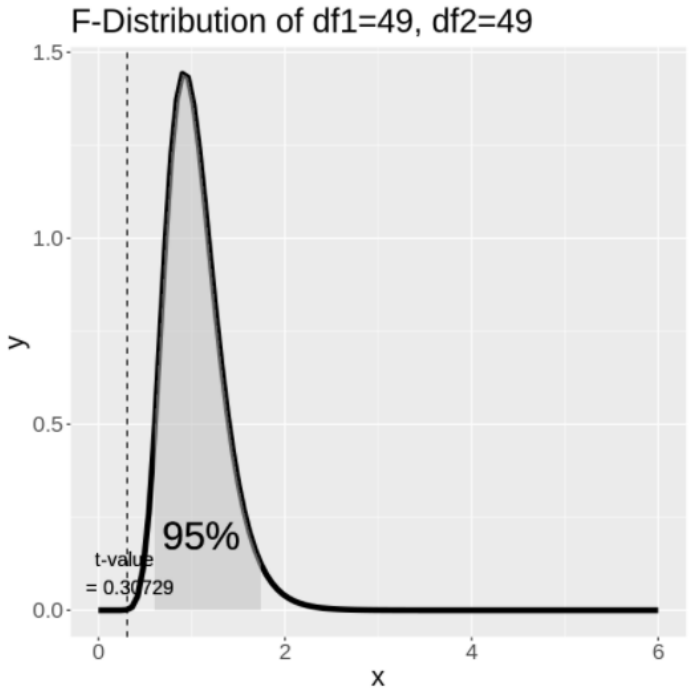
** 두 변수가 독립이면 상관계수가 반드시 0이지만, 상관계수가 0이라고 두 변수가 반드시 독립인 것은 아니다.

F-test

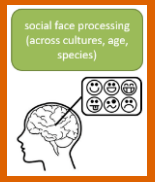
F 분포를 이용한 검정으로, 주로 분산에 대한 분석을 할 때 이용된다.

ex) Iris 꽃의 'setosa'종과 'virginica' 종의 꽃받침 길이의 분산이 같은가?

```
### F-test
var(iris[iris$Species == 'setosa','Sepal.Length'])
var(iris[iris$Species == 'virginica','Sepal.Length'])
var.test(Sepal.Length~Species, data = iris_2group)
```



ANOVA



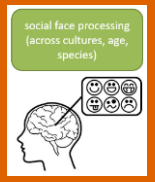
CognitiveSystems

분산분석(Analysis Of Variance: ANOVA)
비교해야될 대상이 **3개 이상**일 때, 해당 표본들의 평균이 차이가 있는지 검정.
(두 표본의 ANOVA 분석 = T-test)

ex) Iris의 종류에 따라 꽃받침 길이가 다른가?

```
### ANOVA
result_aov = aov(Sepal.Length ~ Species, data = iris)
summary(result_aov)
```

비모수적 검정



CognitiveSystems

모수적 검정: 모수에 대한 가정을 하고, 이에 대한 검정 통계량을 이용해 검정을 실시

ex) T-test, F-test, ANOVA 등

비모수적 검정: 자료가 추출된 모집단의 분포에 대해 아무 제약을 가하지 않고 검정을 실시하는 방법. 관측된 자료가 특정 분포를 따른다고 가정할 수 없는 경우에 이용. 자료의 수가 많지 않거나 개체 간 서열관계를 나타낼 경우 사용

ex) KNN, 순위상관계수 등

3. 회귀분석

상관관계 vs 인과관계

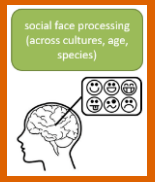
상관관계: 두 변수 간의 연관된 정도를 설명

인과관계: 두 변수 간의 과정과 결과에 대한 설명

ex)

비가 내리는 것과 껌 판매량은 상관관계가 있다. (O)
비가 자주 내릴수록 껌 판매량이 높아진다. (X)

회귀분석



CognitiveSystems

하나나 그 이상의 독립변수(X)들이 종속변수(Y)에 미치는 영향을 추정할 수 있는 통계기법.

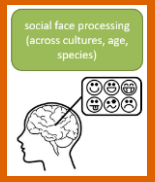
$$Y = f(X) + \epsilon$$

다음과 같이 Y를 X에 대한 함수 (f)와 오차(ϵ)로 정의를 하고, 오차를 최소화 할 수 있는 함수의 계수를 추정한다. 이렇게 추정된 회귀선은 흩어진 점들에 대해 가장 적합한 선으로 여겨진다.

1) 선형회귀분석(Linear Regression): 실수형 Y 변수를 추정

2) 로지스틱 회귀분석(Logistic Regression): 범주형 Y 변수를 추정 (True, False)

회귀분석 Process



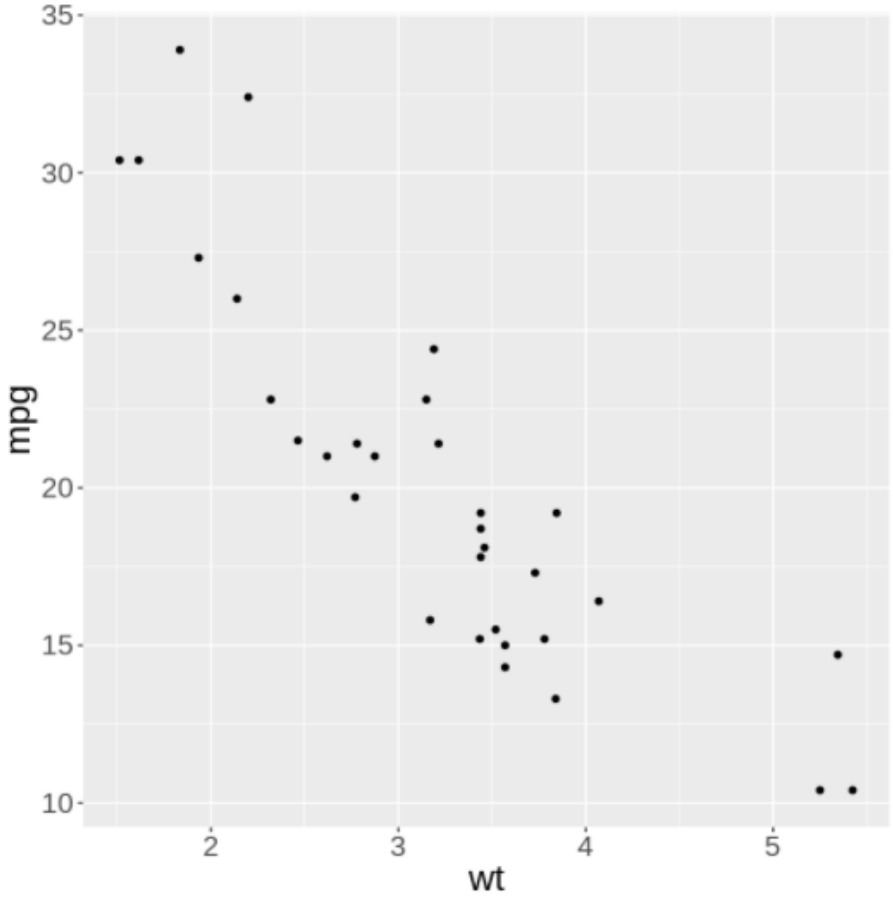
CognitiveSystems

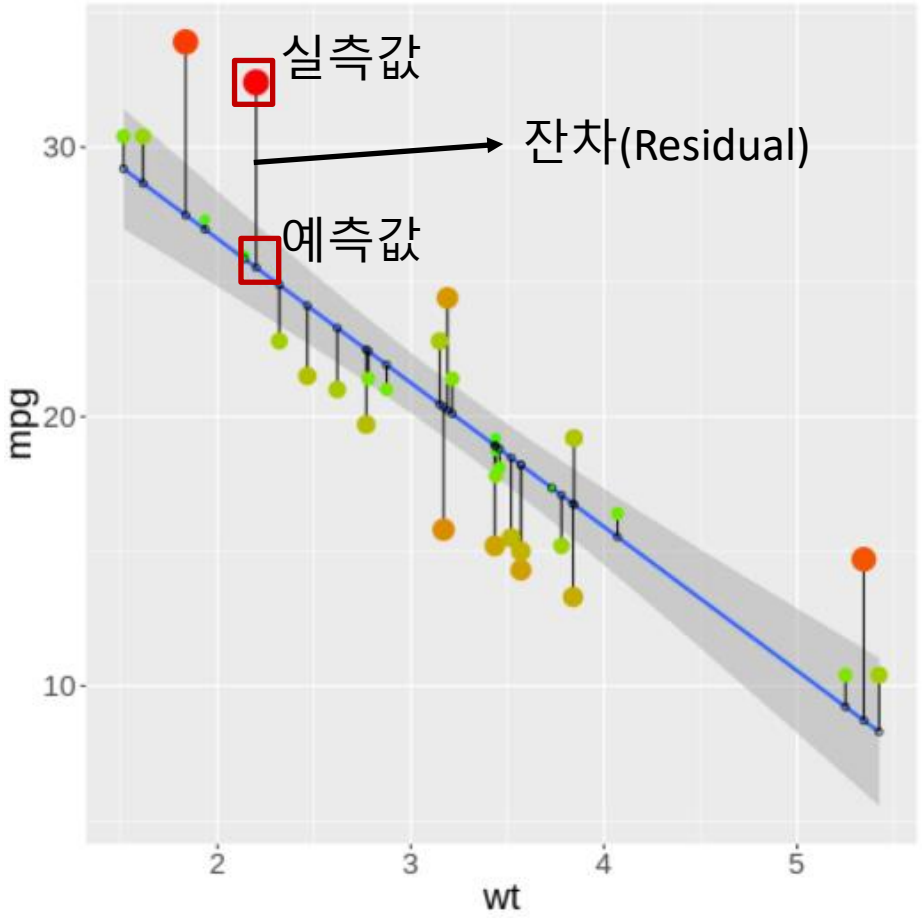
1. 설정한 모델이 통계적으로 **유의미**한가?
2. 모델이 얼마나 **설명력**을 갖는가?
3. 어떤 독립변수가 종속변수의 **변화에 기여**하는가?
4. 각 독립변수들이 종속변수를 **얼마나 변화**시키는가?
5. 모형이 데이터를 잘 **적합**하고 있는가?

단순선형회귀

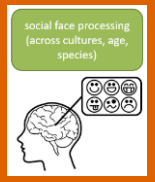
1개의 독립변수와 종속변수의 관계를 직선으로 가정한 회귀분석

가정: $y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, 도출: $\hat{y_i} = b_0 + b_1 x_i$





단순선형회귀 코드



CognitiveSystems

다음 예시는 cars데이터에 대한 회귀분석이다. 우리는 다음과 같은 모델을 구성하고 싶다.

$$Distance_i = \beta_0 + \beta_1 Speed_i + \epsilon_i$$

```
> lm(dist ~ speed, data = cars)

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932

> summary(lm(dist ~ speed, data = cars))

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

단순 결과 코드

lm(y ~ x, data = data.frame)

lm: Linear model

해석 코드:

summary(linear model)

단순선형회귀 해석

CognitiveSystems

1. 설정한 모델이 통계적으로 **유의미**한가?

```
> summary(lm(dist ~ speed, data = cars))

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$$F(1,48) = 89.57$$

$$\text{P-value} = 1.49\text{e-}12 < 0.01$$

따라서 유의수준 99%에서 귀무가설을 기각하고 대립가설을 채택
모델은 통계적으로 **유의미하다**.

단순선형회귀 해석

CognitiveSystems

2. 모델이 얼마나 **설명력**을 갖는가?

```
> summary(lm(dist ~ speed, data = cars))

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$$R^2 = 0.6511$$

본 모델은 전체 데이터의 65%정도 설명 가능한 모델이다.

단순선형회귀 해석

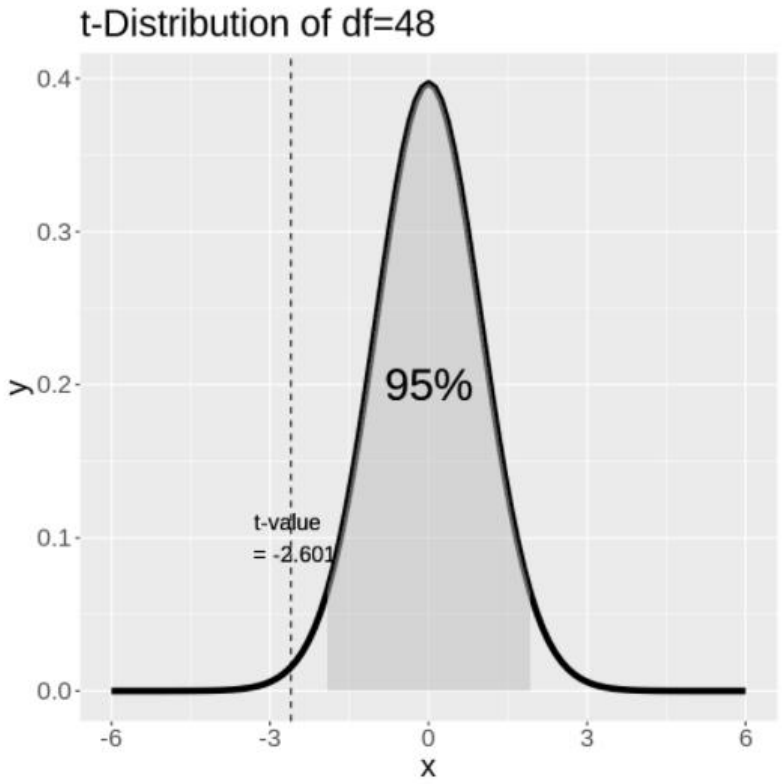
3. 어떤 독립변수가 종속변수의 **변화에 기여**하는가?

b_0, b_1 의 가설검정:

해당 독립변수와 상수항이 종속변수의 변화에 영향을 미치는가?

$$H_0: b_0 = 0 \quad \text{vs} \quad H_1: b_0 \neq 0, \qquad H_0: b_1 = 0 \quad \text{vs} \quad H_1: b_1 \neq 0$$

```
> summary(lm(dist ~ speed, data = cars))

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```



단순선형회귀 해석

CognitiveSystems

4. 각 독립변수들이 종속변수를 **얼마나 변화**시키는가?

```
> summary(lm(dist ~ speed, data = cars))

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$$b_0 = -17.5791, b_1 = 3.9324$$

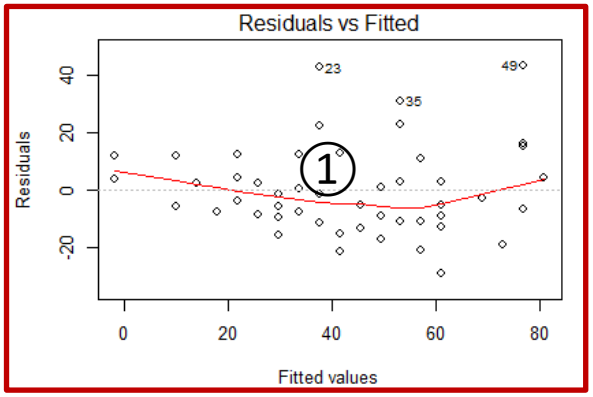
따라서 우리의 모델은 다음과 같이 설명할 수 있다.

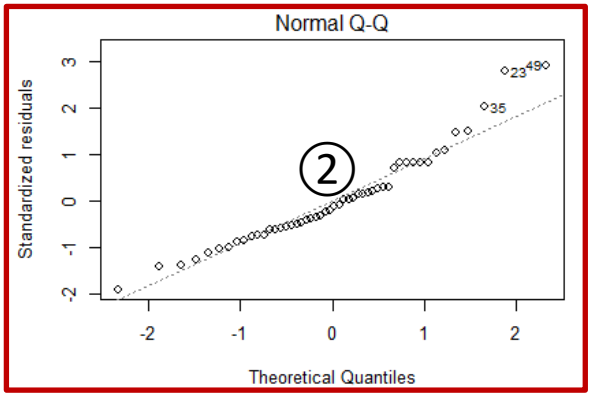
$$Distance_i = -17.5791 + 3.9324 \times Speed_i + \epsilon_i$$

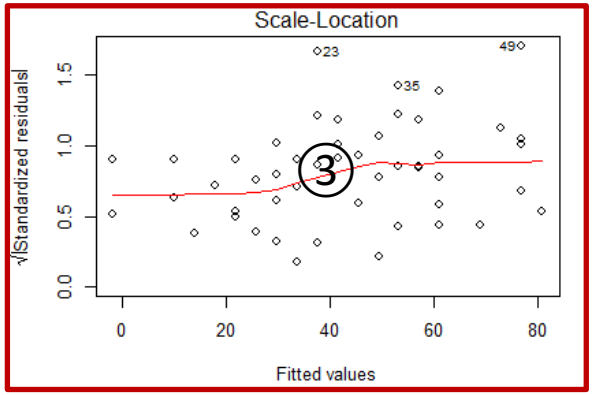
단순선형회귀 해석

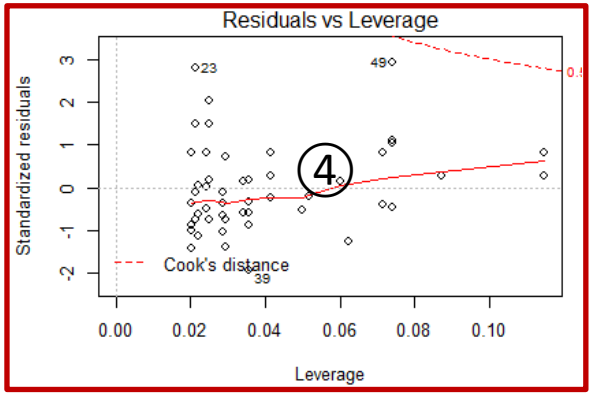
5. 모형이 데이터를 잘 **적합**하고 있는가? > plot(linear model)

① 산출된 수치와 잔차를 시각화. 0을 중심으로 고르게 분포해야 적합한 모델.
② 잔차가 표준편차를 따라 분포하는지 도식화
③ 잔차를 표준화하여 루트 적용. 트랜드가 없어야 적합
④각 잔차의 영향을 보여줌.

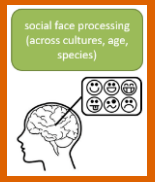








다중선형회귀



CognitiveSystems

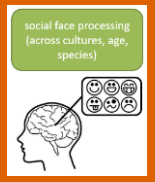
2개 이상의 독립변수와 종속변수의 관계를 선형으로 가정한 회귀분석

가정: $y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + \epsilon_i$
도출: $\hat{y_i} = b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}$

** 다중선형회귀의 특징 및 주의사항:

1) 어떤 변수라도 추가되면 R^2는 항상 높아진다.
2) 어떤 변수와 같이 회귀분석을 하느냐에 따라 해당 변수가 유의미 할수도, 그렇지 않을 수도 있다. – 변수 간의 상관성을 고려
3) 다중 공선성(Multicollinearity, 변수 사이에 완벽한 선형관계)
   ex) X3 = (X1+X2)/2라는 변수를 추가하면, X1,X2,X3 사이에 다중 공선성 발생.

다중선형회귀 코드



CognitiveSystems

다음 예시는 mtcars데이터에 대한 회귀분석이다. 우리는 다음과 같은 모델을 구성하고 싶다.

$$Disp_i = \beta_0 + \beta_1 cyl_i + \beta_2 hp_i + \beta_3 wt_i + \epsilon_i$$

```
> lm(disp~cyl+hp+wt, data=mtcars)

Call:
lm(formula = disp ~ cyl + hp + wt, data = mtcars)

Coefficients:
(Intercept)          cyl           hp           wt
   -179.0419      30.3212       0.2156      59.2220

> summary(lm(disp~cyl+hp+wt, data=mtcars))

Call:
lm(formula = disp ~ cyl + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-67.561 -25.534   6.211  29.607  71.041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -179.0419    28.8072  -6.215 1.03e-06 ***
cyl           30.3212     8.8817   3.414  0.00197 **
hp             0.2156     0.1915   1.126  0.26980
wt            59.2220    11.9393   4.960 3.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.49 on 28 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.8933
F-statistic: 87.48 on 3 and 28 DF,  p-value: 2.475e-14
```

단순 결과 코드

lm(y ~ x1+…+xk, data =data)

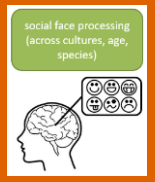
모든 변수 사용:

lm(y ~ ., data = data)

해석 코드:

summary(linear model)

다중선형회귀 해석



CognitiveSystems

1. 설정한 모델이 통계적으로 **유의미**한가?

```
> summary(lm(disp~cyl+hp+wt, data=mtcars))

Call:
lm(formula = disp ~ cyl + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-67.561 -25.534   6.211  29.607  71.041

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -179.0419    28.8072  -6.215 1.03e-06 ***
cyl           30.3212     8.8817   3.414  0.00197 **
hp             0.2156     0.1915   1.126  0.26980
wt            59.2220    11.9393   4.960 3.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.49 on 28 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.8933
F-statistic: 87.48 on 3 and 28 DF,  p-value: 2.475e-14
```

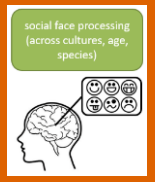
$$F(3,28) = 87.48$$

$$\text{P-value} = 2.4575\text{e-}14 < 0.01$$

따라서 유의수준 99%에서 귀무가설을 기각하고 대립가설을 채택

모델은 통계적으로 **유의미하다**.

다중선형회귀 해석



CognitiveSystems

2. 모델이 얼마나 **설명력**을 갖는가?

```
> summary(lm(disp~cyl+hp+wt, data=mtcars))

Call:
lm(formula = disp ~ cyl + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-67.561 -25.534   6.211  29.607  71.041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -179.0419    28.8072  -6.215 1.03e-06 ***
cyl           30.3212     8.8817   3.414  0.00197 **
hp             0.2156     0.1915   1.126  0.26980
wt            59.2220    11.9393   4.960 3.09e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.49 on 28 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.8933
F-statistic: 87.48 on 3 and 28 DF,  p-value: 2.475e-14
```

$$R^2 = 0.9036$$

본 모델은 전체 데이터의 90.36% 정도 설명 가능한 모델이다.

다중선형회귀 해석

CognitiveSystems

3. 어떤 독립변수가 종속변수의 **변화에 기여**하는가?

b_0, b_1 의 가설검정:

해당 독립변수와 상수항이 종속변수의 변화에 영향을 미치는가?

```
> summary(lm(disp~cyl+hp+wt, data=mtcars))

Call:
lm(formula = disp ~ cyl + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-67.561 -25.534   6.211  29.607  71.041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -179.0419    28.8072  -6.215 1.03e-06 ***
cyl           30.3212     8.8817   3.414  0.00197 **
hp             0.2156     0.1915   1.126  0.26980
wt            59.2220    11.9393   4.960 3.09e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.49 on 28 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.8933
F-statistic: 87.48 on 3 and 28 DF,  p-value: 2.475e-14
```

** hp의 경우, cyl와 wt가 있을 경우 disp에 대한 영향이 유의미하다고 할 수 없다.

**** 현재 모델에서 유의미하지 않다고 모든 상황에서 유의미하지 않다는 것은 아님을 주의! ****

다중선형회귀 해석

CognitiveSystems

4. 각 독립변수들이 종속변수를 **얼마나 변화**시키는가?

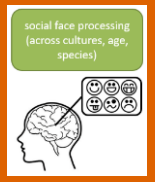
```
> summary(lm(disp~cyl+hp+wt, data=mtcars))

Call:
lm(formula = disp ~ cyl + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-67.561 -25.534   6.211  29.607  71.041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -179.0419    28.8072  -6.215 1.03e-06 ***
cyl           30.3212     8.8817   3.414  0.00197 **
hp             0.2156     0.1915   1.126  0.26980
wt            59.2220    11.9393   4.960 3.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.49 on 28 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.8933
F-statistic: 87.48 on 3 and 28 DF,  p-value: 2.475e-14
```

$b_0 = -179.0419, b_1 = 30.3212$
$b_2 = 0.2156, b_3 = 59.2220$

따라서 우리의 모델은 다음과 같이 설명할 수 있다.

$$Disp_i = -179.0419 + 30.3212 \times cyl_i + 0.2156 \times hp_i + 59.2220 \times wt_i + \epsilon_i$$

변수선택



CognitiveSystems

다중회귀분석에서 유의미한 변수만을 선택하는 방법

RSS(Residual Sum of Sqaure), AIC(Akaike information criterion) 등의 정보를 이용하여 변수를 선별한다.

1. 전진선택법(Forward selection): 절편만 있는 상수모형에서 차례로 변수를 추가하는 선택법

```
#Forward selection
step(lm(disp~1,data=mtcars),direction="forward", scope=(~cyl+hp+wt))
```

2. 후진선택법(Backward selection): 모든 변수가 있는 모형에서 차례로 변수를 제거하는 선택법

```
#Backward selection
step(lm(disp~cyl+hp+wt,data=mtcars),direction="backward")
```

3. 혼합선택법(Mixed/Stepwise selection): 전진선택법+후진선택법

```
#Mixed selection
step(lm(disp~1,data=mtcars),direction="both", scope=(~cyl+hp+wt))
```

로지스틱 회귀

종속변수가 범주형(0:1, True:False)일때의 관계를 로지스틱 모형으로 가정한 회귀분석

$$p(y_i = 1|X) = p(X) = \frac{\exp(b_0+b_1x_{1,i}+\cdots+b_kx_{k,i})}{1+\exp(b_0+b_1x_{1,i}+\cdots+b_kx_{k,i})}$$

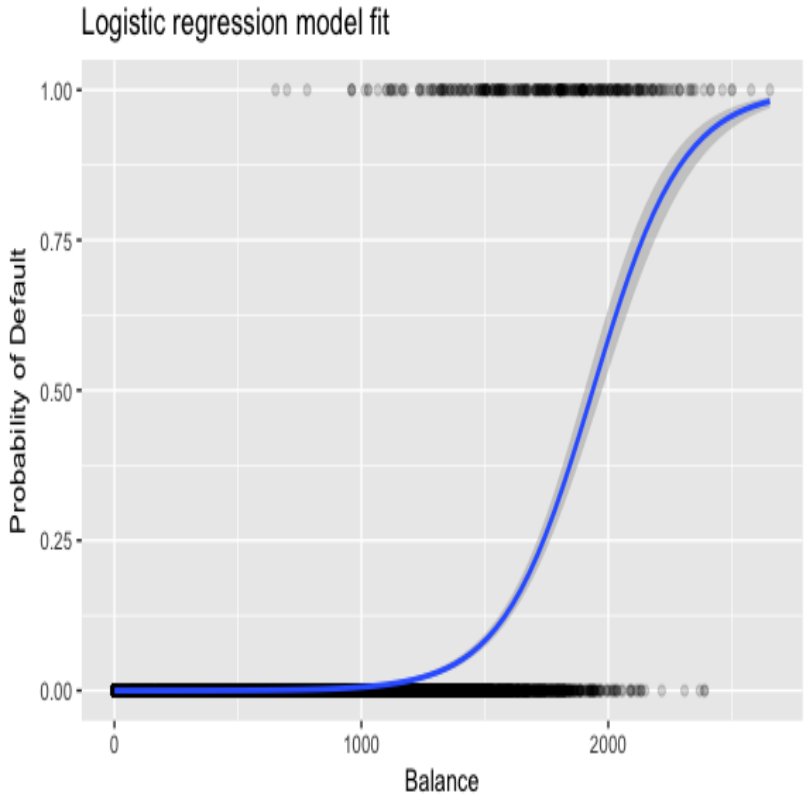
$$p(y_i = 0|X) = 1 - p(X)$$

$$\frac{p(X)}{1-p(X)} = \exp(b_0 + b_1x_{1,i} + \cdots + b_kx_{k,i})$$

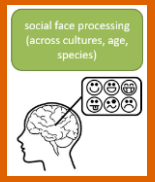
$$\log(\frac{p(X)}{1-p(X)}) = b_0 + b_1x_{1,i} + \cdots + b_kx_{k,i}$$

* $\log(\frac{p(X)}{1-p(X)})$: Logit(0~∞)

True가 일어날 확률이 작아질 수록 0에 가까워지고 반대로 True가 일어날 확률이 높아질수록 ∞에 가까워진다.



로지스틱 회귀분석 코드



CognitiveSystems

다음 예시는 MASS 패키지의 biops데이터에 대한 로지스틱 회귀분석이다.
우리는 다음과 같은 모델을 구성하고 싶다.

$$\log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 V1_i + \beta_2 V2_i + \beta_3 V4_i + \beta_4 V6_i + \beta_5 V7_i + \epsilon_i$$

```
> summary(glm(class ~ V1+V2+V4+V6+V7, data = prep_biopsy, family = binomial))

Call:
glm(formula = class ~ V1 + V2 + V4 + V6 + V7, family = binomial,
    data = prep_biopsy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3504  -0.1318  -0.0644   0.0244   2.2969

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5850     1.0455  -9.168  < 2e-16 ***
V1            0.6555     0.1344   4.878 1.07e-06 ***
V2            0.3979     0.1362   2.922  0.00348 **
V4            0.3370     0.1176   2.865  0.00418 **
V6            0.4207     0.0903   4.659 3.18e-06 ***
V7            0.5326     0.1645   3.238  0.00120 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 114.60  on 677  degrees of freedom
AIC: 126.6

Number of Fisher Scoring iterations: 8
```

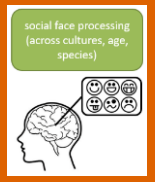
단순 결과 코드

glm(y ~ x, data = data, +
family = 'binomial')

해석 코드:

summary(glm result)

로지스틱 회귀분석 해석



CognitiveSystems

1. 모델이 얼마나 **설명력**을 갖는가?

```
logit.probs = predict(result3,prep_biopsy, type = "response")
logit.pred = ifelse(logit.probs > .5, 'malignant','benign')
table(logit.pred, prep_biopsy$class)
```

```
logit.pred  benign malignant
  benign       433        12
  malignant     11       227
```

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

$$precision = \frac{TP}{TP + FP}$$

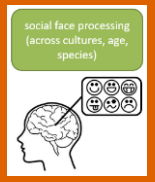
$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

로지스틱 회귀분석 해석



CognitiveSystems

2. 어떤 독립변수가 종속변수의 **변화에 기여**하는가?

```
> summary(glm(class ~ V1+V2+V4+V6+V7, data = prep_biopsy, family = binomial))

Call:
glm(formula = class ~ V1 + V2 + V4 + V6 + V7, family = binomial,
    data = prep_biopsy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3504  -0.1318  -0.0644   0.0244   2.2969

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5850     1.0455  -9.168  < 2e-16 ***
V1            0.6555     0.1344   4.878 1.07e-06 ***
V2            0.3979     0.1362   2.922  0.00348 **
V4            0.3370     0.1176   2.865  0.00418 **
V6            0.4207     0.0903   4.659 3.18e-06 ***
V7            0.5326     0.1645   3.238  0.00120 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 114.60  on 677  degrees of freedom
AIC: 126.6

Number of Fisher Scoring iterations: 8
```

* 모든 변수가 종속변수와 유의미한 관계가 있음.

로지스틱 회귀분석 해석

CognitiveSystems

3. 각 독립변수들이 종속변수를 **얼마나 변화**시키는가?

```
> summary(glm(class ~ V1+V2+V4+V6+V7, data = prep_biopsy, family = binomial))

Call:
glm(formula = class ~ V1 + V2 + V4 + V6 + V7, family = binomial,
    data = prep_biopsy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3504  -0.1318  -0.0644   0.0244   2.2969

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.5850     1.0455  -9.168  < 2e-16 ***
V1           0.6555     0.1344   4.878 1.07e-06 ***
V2           0.3979     0.1362   2.922  0.00348 **
V4           0.3370     0.1176   2.865  0.00418 **
V6           0.4207     0.0903   4.659 3.18e-06 ***
V7           0.5326     0.1645   3.238  0.00120 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 114.60  on 677  degrees of freedom
AIC: 126.6

Number of Fisher Scoring iterations: 8
```

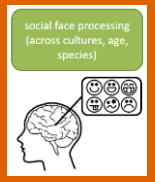
$$b_0 = -9.5850, b_1 = 0.6555$$
$$b_2 = 0.3979,\ b_3 = 0.3370$$
$$b_4 = 0.4207, b_5 = 0.5326$$

따라서 우리의 모델은 다음과 같이 설명할 수 있다.

$$\log\left(\frac{P(X)}{1-P(X)}\right) = -9.5850 + 0.6555 * V1_i + 0.3979 * V2_i + 0.3370 * V4_i + 0.4207 * V6_i + 0.5326 * V7_i + \epsilon_i$$

실습

실습



CognitiveSystems

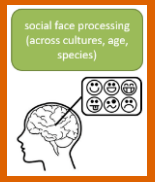
1. 다음은 각 팀에 대한 근속 연수를 조사한 데이터이다.

A팀: 13 8 5 6 9 9 13 5 9 14 6 9 12 12 13
B팀: 3 9 10 20 5 6 4 1 19 17 2 4 6 11 18

위 데이터를 데이터 프레임을 생성하여 각 팀 근속 연수의 평균과 분산을 계산하고, 두 팀의 근속 연수가 차이가 나는지에 대해 유의수준 5% 하에서 검정하여라.

* F검정을 통해 두 집단의 분산이 같은지를 먼저 검정하고, T-test의 var.equal이 True, False인지 판별하여 적용하여라.

실습



CognitiveSystems

2. 다음은 어떤 앱에서 온라인으로 의류를 사는 소비자들의 정보에 대한 데이터이다. 해당 앱은 소비자들이 원하는 옷을 상담해주고, 그에 따라 앱이나 웹사이트에서 의류를 선택 해 소비자들에게 추천하는 서비스를 제공하고 있다. 각 데이터의 설명은 다음과 같다.

-Avg. Session Length: 평균적으로 얼마나 상담을 받는지

-Time on App: 앱 사용 시간

-Time on Website: 웹사이트 사용 시간

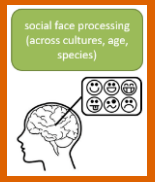
-Length of Membership: 멤버쉽 소지 기간

-Yearly Amount Spent: 연 평균 소비

어떤 변수가 연 평균 소비에 영향을 주는지 설명하여라. 그리고 소비자들의 연 평균 소비를 증진하기 위해서는 어떤 방안을 제시하는 것이 좋을 지 분석한 내용을 기반으로 제안해 보아라.

출처: https://www.kaggle.com/kolawale/focusing-on-mobile-app-or-website

실습



CognitiveSystems

3. 다음은 심장 질환에 대한 연구로, 각 환자들마다 다음의 정보를 포함하고 있는 데이터이다.

- age: 나이
- currentSmoker: 현재 흡연을 하고 있는지
- BPMeds: 혈압 치료를 받고 있는지
- diabetes: 당뇨가 있는지
- totChol: 총 콜레스테롤
- sysBP: 혈압
- BMI: BMI
- heartrate : 심박수
- glucose : 글루코오스
- TenYearCHD: 10년 안에 심장병이 발현했는지

어떤 변수가 심장병 발병에 영향을 주는지 설명하여라. 그리고 심장병 발병을 예방하기 위해 어떤 방안이 좋을 지 분석한 내용을 기반으로 제안해 보아라.

출처: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset