Part-IV. 예측 분석(기계학습 알고리즘)

15. 회귀분석
16. 분류분석
17. 군집분석
18. 연관분석

기계학습 분류

1. 지도학습(Supervised Learning)

- 인간 개입에 의한 분석 방법
- 종속변수(y) 존재 : 입력 데이터에 정답 포함
- 분석방법 : 가설검정(확률/통계) → 인문.사회.심리 계열(300년)
- 분석유형 : 회귀분석, 분류분석, 시계열 분석 → 추론통계 기반

2. 비지도학습(unSupervised Learning)

- 컴퓨터 기계학습에 의한 분석 방법
- 종속변수(y) 없음 : 입력 데이터에 정답 없음
- 분석방법 : 규칙(패턴분석) → 공학.자연과학 계열(100년)
- 분석유형 : 연관분석, 군집분석 → 데이터마이닝 기반

1. 학습방식에 따른 분류

● 지도학습과 비지도학습

분류	지도학습	비지도학습
주 관	사람의 개입에 의한 학습	컴퓨터에 의한 기계학습
기 법	확률과 통계 기반 추론통계	패턴분석 기반 데이터 마이닝
유 형	회귀분석, 분류분석(y변수 있음)	군집분석, 연관분석(y변수 없음)
분 야	인문, 사회 계열	공학, 자연 계열

1. 학습방식에 따른 분류

1. 회귀분석 : 인과관계 예측(수치예측)

2. 분류분석 : 고객 이탈분석(번호이동, 반응고객 대상 정보 제공)

3. 군집분석 : 그룹화를 통한 예측(그룹 특성 차이 분석-고객집단 이해)

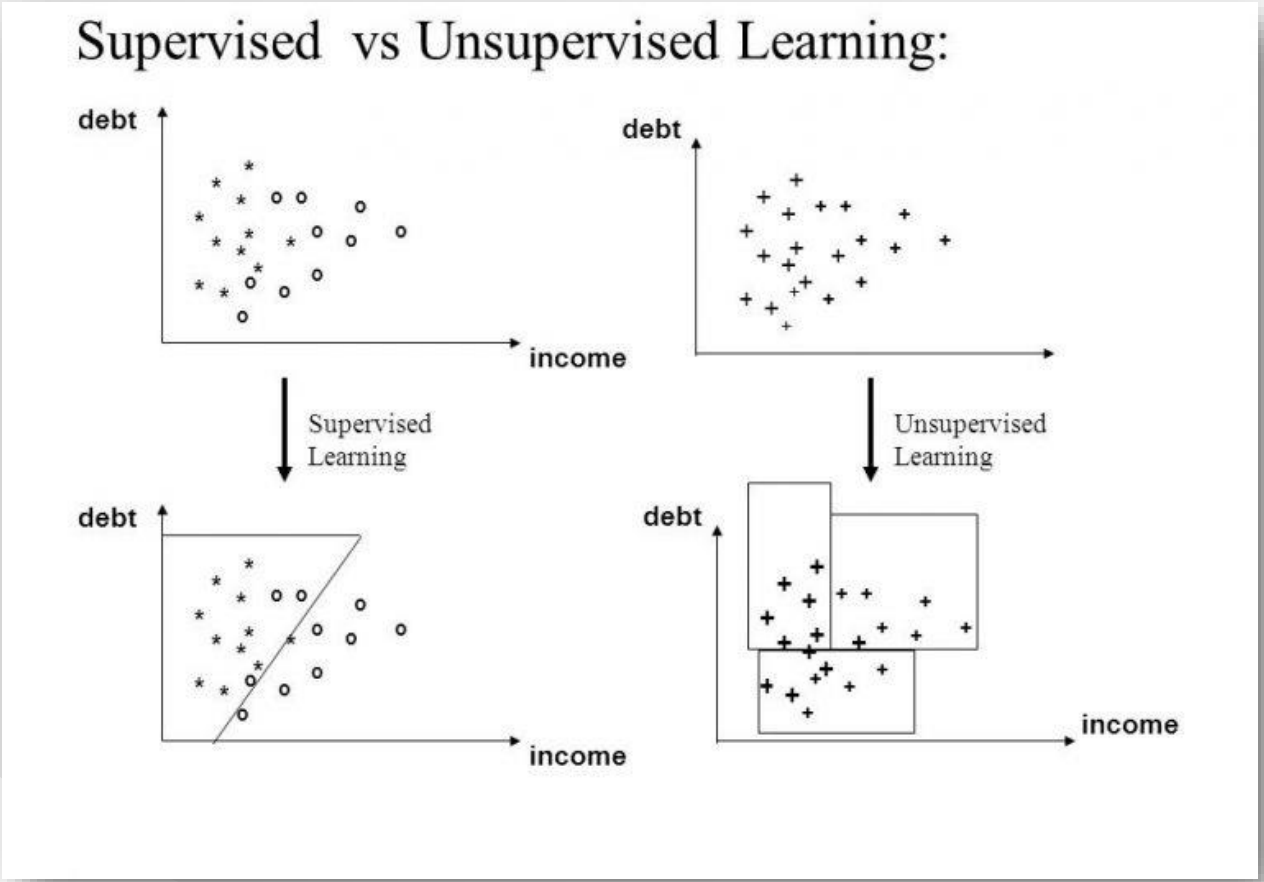
4. 연관분석 : 상품구매 규칙을 통한 구매 패턴 예측(상품 연관성)

❖ 분류(Classification) vs 군집(Clustering) 분석
분류 분석은 이미 각 계급(클러스터)이 어떻게 정의 되는지 알고 있음

1. 학습방식에 따른 분류

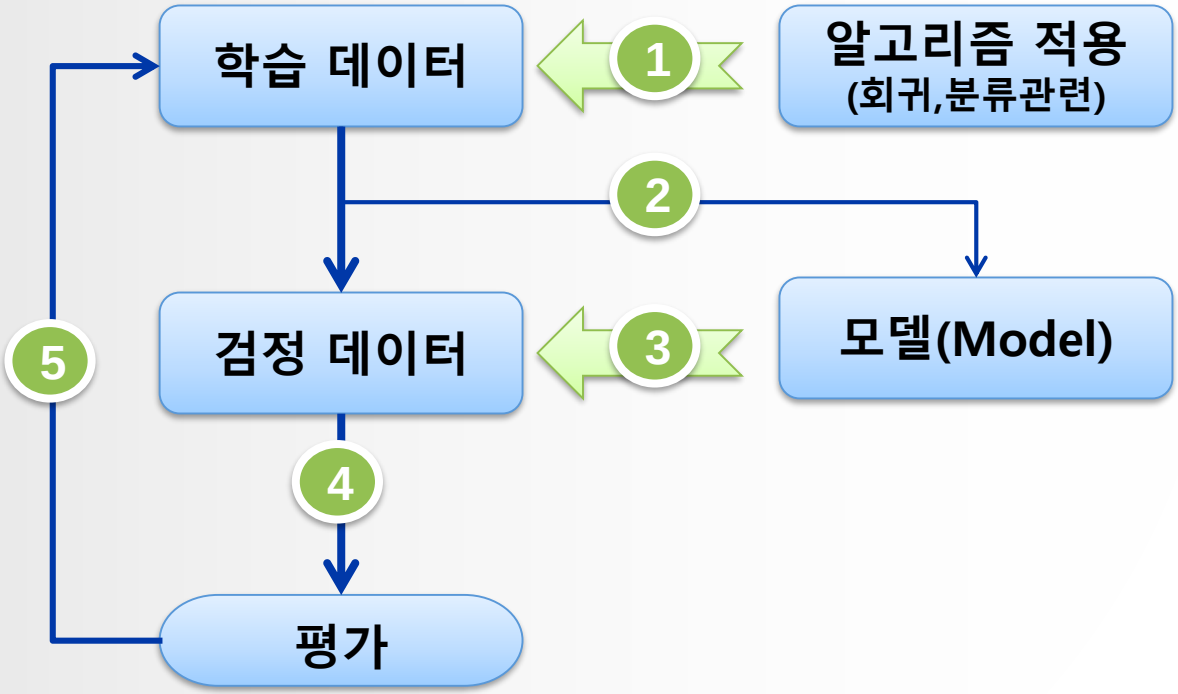
❖ 분류(Classification) vs 군집(Clustering) 분석

분류 분석은 이미 각 계급(클러스터)이 어떻게 정의 되는지 알고 있음(y 존재)



지도학습

- 지도학습(Supervised Learning) 절차



15-1. 선형 회귀분석

Chap15_1_LinearRegration 수업내용

1. 회귀분석 개요
2. 단순회귀분석
3. 다중회귀분석(다중공선성 문제)
4. 효과적인 변수 선택법
5. 기계학습

1. 회귀분석 개요

- **회귀분석(Regression Analysis)**
  - 특정 변수(독립변수)가 다른 변수(종속변수)에 어떠한 영향을 미치는가 (인과관계 분석)
  - 예) 가격은 제품 만족도에 영향을 미치는가?
  - 한 변수의 값으로 다른 변수의 값 예언

[참고] 인과관계(因果關係) : 변수A가 변수B의 값이 변하는 원인이 되는 관계(변수A : 독립변수, 변수B : 종속변수)

- 상관관계분석 : 변수 간의 관련성 분석
- 회귀분석 : 변수 간의 인과관계 분석

1. 회귀분석 개요

- **상관관계 vs 인과관계**
  - 상관관계가 높다고 반드시 인과관계가 있다고 볼 수 없음
  - 상관관계 : 붉은포도주 → 심장병 발병률
    - ✓ 포도주는 심장병에 효과가 있지만, 포도주 양을 늘리거나, 줄일 때 심장병 발병률이 줄어들거나, 높아지는 것은 아니다.
  - 인과관계 : 스트레스 → 심장병 발병률
    - ✓ 스트레스로 인한 긴장과 분노는 심장 박동수나 강도를 높이고, 심장의 산소 소비량을 증가, 관상동맥은 수축되어 혈액순환이 적절하지 못하여 심장에 영향을 미친다.

1. 회귀분석 개요

【회귀분석 중요사항】

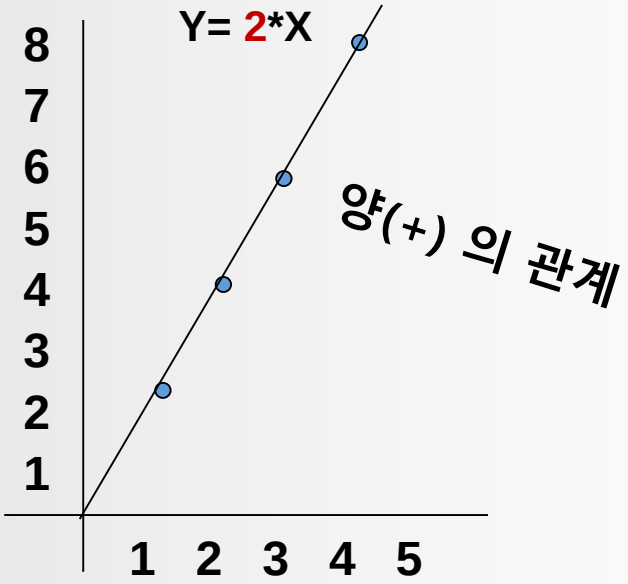
- '통계분석의 **꽃**' ➔ 가장 강력하고, 많이 이용
- 종속변수에 영향을 미치는 변수를 규명(변수 선형 관계 분석)
- 독립변수와 종속변수의 관련성 강도
- 독립변수의 변화에 따른 종속변수 변화 예측
- **회귀 방정식**($Y=\alpha+\beta X$) : 회귀선 추정
  - ✓ Y:종속변수, α :상수, β:회귀계수, X:독립변수
- 독립변수와 종속변수가 모두 <u>등간척도 또는 비율척도</u> 구성

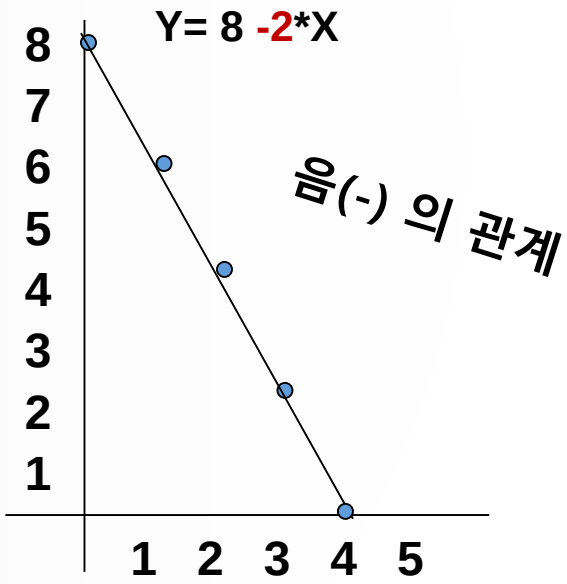
1. 회귀분석 개요

- **선형 회귀 방정식(1차 함수) : 회귀선 추정**

  Y= a*X + b

  (Y : 종속변수, a : 기울기, X : 독립변수, b : 절편)





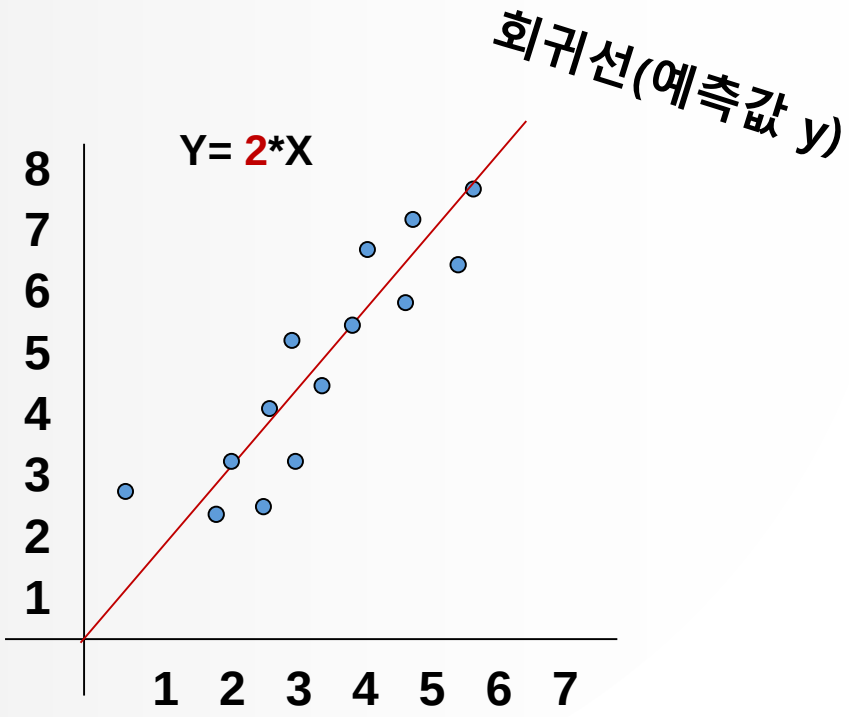
  - 회귀방정식에 의해서 x가 10일 때 y는 20 예측 -> 회귀분석은 **수치 예측**

1. 회귀분석 개요

- **최소자승법 적용 회귀선**

  회귀방정식에 의해서 그려진 y의 추세선

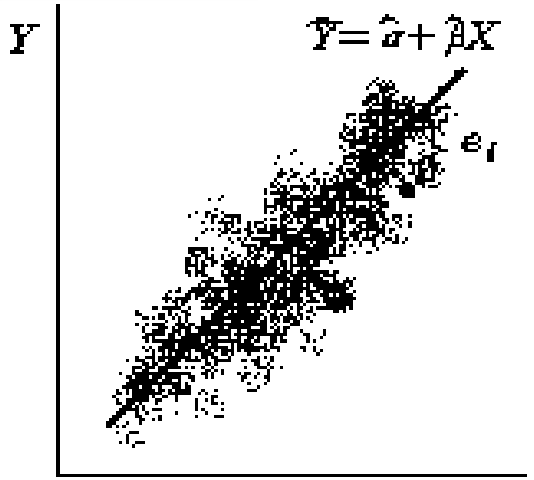
  산포도 각 점의 위치를 기준으로 정중앙 통과하는 회귀선 추정 방법



1. 회귀분석 개요

【회귀방정식】

- 회귀 방정식 -> 회귀선 추정
  - ✓ Y=a+βX : Y:종속변수, a:상수, β:회귀계수, X:독립변수
- 회귀계수(β) : 단위시간에 따라 변하는 양(기울기)이며, 회귀선을 추정함에 있어 최소자승법 이용
- 최소자승법 : 산포도에 위치한 각 점에서 회귀선에 수직으로 이르는 값의 제곱의 합 최소가 되는 선(정중앙을 통과하는 직선)을 최적의 회귀선으로 추정



산포도와 회귀선

1. 회귀분석 개요

【회귀분석의 기본 가정 충족】

- 선형성 : 독립변수와 종속변수가 선형적 이어야한다. 【회귀선 확인】
- 오차의 정규성 : 오차란 종속변수의 관측값과 예측값 간의 차이를 말하며, 오차의 기대값은 0이며, 정규분포를 이루어야한다. 【정규성 검정 확인】
- 오차의 독립성 : 오차들은 서로 독립적 이어야한다.【더빈-왓슨 값 확인 】
- 오차의 등분산성 : 오차들의 분산이 일정해야한다. 【표준잔차와 표준예측치 도표】
- 다중공선성 : 다중 회귀분석을 수행할 경우 3개 이상의 독립변수들 간의 강한 상관관계로 인한 문제가 발생되지 않아야 한다.【분산팽창요인(VIF) 확인】 ※ 회귀분석을 수행하기 위해서는 위와 같은 기본 가정이 충족되어야 분석이 가능하며, 이러한 기본 가정을 토대로 일반적인 회귀분석을 위한 절차는 다음과 같다.

2. 단순 회귀분석

- **단순 회귀분석**
  - 독립변수와 종속변수 각각 1개
  - 독립변수가 종속변수에 미치는 인과관계 분석

【단순 회귀분석 가설】

음료수 제품의 당도와 가격수준을 결정하는 제품적절성(독립변수)은 제품만족도(종속변수)에 **정(+)**의 영향을 미칠 것이다.

2. 단순 회귀분석

```
# 단순회귀분석
 형식) lm(formula= y ~ x 변수, data)
  # x:독립, y:종속, data=data.frame
  # lm() 함수 -> x변수를 대상으로 y변수 값 유추

str(result)
y = result$만족도 # 종속변수
x = result$적절성 # 독립변수
result.lm <- lm(formula=y ~ x, data=result)

# 단순선형회귀 분석 결과 보기
summary(result.lm)
```

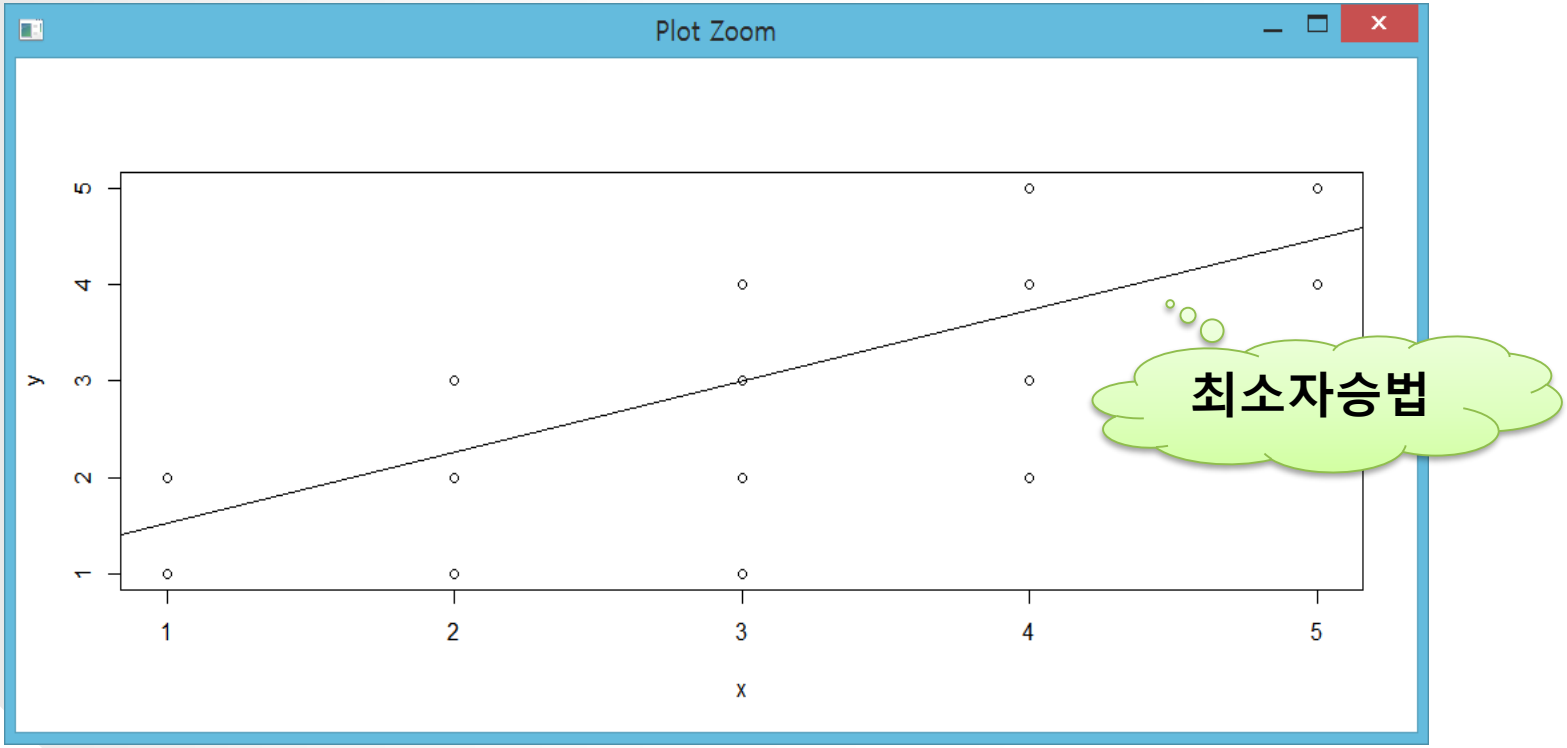
2) 단순 회귀분석

```
summary(result.lm)
#Coefficients: 계수
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)  0.77886    0.12416   6.273 1.45e-09 ***
#x            0.73928    0.03823  19.340  < 2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 0.5329 on 262 degrees of freedom
#Multiple R-squared:  0.5881,  Adjusted R-squared:  0.5865
#F-statistic:   374 on 1 and 262 DF,  p-value: < 2.2e-16
################<회귀분석 결과 해석>#########################
# 결정계수(Coefficients) : R-squared -> 0 ~ 1 사이의 값을 갖는다.
# Multiple R-squared:  0.5881:  독립변수에 의해서 종속변수가 얼마만큼 설명되었는가?
# 설명력 -> 상관(결정)계수 : 58.8% 설명력
# 1에 가까울 수록 설명변수(독립변수)가 설명을 잘한다고 판단
# 모형의 변수 선정이 우수하다는 의미.
# Adjusted R-squared:  0.5865 : 조정된 R값(오차를 감안한 값)<- 이것으로 분석
```

2. 단순 회귀분석

- **회귀방정식에 의해서 회귀선 시각화**
  - ✓ X,Y가 선형 관계를 나타냄



2. 단순 회귀분석

【논문에서 단순 회귀분석 결과 제시 방법】

종속변수	독립변수	표준오차 (Std.Error)	검정통계량(t)	유의확률 (p)
제품만족도	상수	0.124	6.273	1.45e-09 ***
	제품적절성	0.038	19.340	< 2e-16 ***
분석 통계량	Multiple R-squared: 0.5881, Adjusted R-squared: 0.5865 F - statistic: 374 on 1 and 262 DF, p-value: < 2.2e-16			

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

분산분석 :
회귀모델 적합성
(유의확률 0.05이상
부적합

2. 단순 회귀분석

【논문에서 단순 회귀분석 결과 제시 방법】

- 음료수 제품의 당도와 가격수준을 결정하는 제품 적절성은 제품 만족도에 정(正)의 영향을 미칠 것이라는 연구가설을 검정한 결과 검정통계량 t=19.340, p=0.05미만으로 통계적 유의수준 하에서 영향을 미치는 것으로 나타났기 때문에 연구가설을 채택한다.
- 회귀모형은 상관계수 R=.767로 두 변수 간에 다소 높은 상관관계를 나타내며, R^2=.587로 제품 적절성 변수가 제품 만족도를 58.7% 설명하고 있다. 또한 회귀모형의 적합성은 F=374.020(p-value : < 2.2e-16)으로 회귀선이 모형에 적합하다고 볼 수 있다.

2. 단순 회귀분석

【단순 회귀분석 결과 정리 및 기술】

▪ 가설 설정	연구가설(H_1) : 음료수 제품의 적절성은 제품 만족도에 **정(正)**의 영향을 미친다.	
	귀무가설(H_0) : 음료수 제품의 적절성은 제품 만족도에 영향을 미치지 않는다.	
1. 회귀식 모델 적합성	1) 유의순준	α = 0.05
	2) 검정통계량	*F* = 374.020
	3) 유의확률	P-value: < 2.2e-16
	4) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 회귀선이 모델에 적합하다고 볼 수 있다.
2. 가설검정	1) 유의수준	α = 0.05
	2) 검정통계량	t = 19.340
	3) 유의확률	*p* = < 2.2e-16
	4) 결과해석	유의수준 0.05에서 연구가설이 채택되었다. 따라서 제품 적절성이 높을 수록 제품 만족도가 높아지는 경향을 보이고 있다.

3. 다중 회귀분석

- **다중 회귀분석**
  - 여러 개 독립변수가 1개의 종속변수에 미치는 영향 분석

【다중 회귀분석 가설】

음료수 제품의 적절성과 친밀도가 높을 수록 제품 만족도(종속변수)도 높아질 것이다.

3. 다중 회귀분석

(1) 적절성 + 친밀도 -> 만족도

```
y <- result$만족도 # 종속변수
x1 <- result$적절성 # 독립변수
x2 <- result$친밀도 # 독립변수
result.lm <- lm(formula= y ~ x1 + x2, data=result)
summary(result.lm)
```

3. 다중 회귀분석

(2) 학습데이터와 검증데이터 분석

```
# 단계1 : 7:3비율 데이터 샘플링
t <- sample(1:nrow(result), 0.7*nrow(result))

# 단계2 : 학습데이터와 검정데이터 생성
train <- result[t, ] # result중 70%
train # 학습데이터
test <- result[-t, ] # result중 나머지 30%
test # 검정 데이터

# 단계3 : 데이터 분석
result.lm <- lm(formula=만족도 ~ 적절성 + 친밀도, data=train)
summary(result.lm) # 학습데이터 분석
result.lm <- lm(formula=만족도 ~ 적절성 + 친밀도, data=test)
summary(result.lm) # 검정데이터 분석
```

3. 다중 회귀분석

3) 다중공선성(Multicolinearity) 문제

- 독립변수 간의 강한 상관관계로 인해서 회귀분석의 결과를 신뢰할 수 없는 현상
- 생년월일과 나이를 독립변수로 갖는 경우
- 해결방안 : 강한 상관관계를 갖는 독립변수 제거

(1) 다중공선성 문제 확인

```
install.packages("car")
library(car)
fit <- lm(formula=Sepal.Length ~ Sepal.Width+Petal.Length+Petal.Width,
    data=train)
vif(fit)
sqrt(vif(fit))>2 # root(VIF)가 2 이상인 것은 다중공선성 문제 의심
```

3. 다중 회귀분석

(2) iris 변수 간의 상관계수 구하기(단,Species제외)

```
cor(iris[,-5])
```

(3) 학습데이터와 검정데이터 분류

```
x <- sample(1:nrow(iris), 0.7*nrow(iris)) # 전체 70% 추출
train <- iris[x, ]
test <- iris[-x, ]
```

(4) Petal.Width 변수를 제거한 후 회귀분석

```
result.lm <- lm(formula=Sepal.Length ~ Sepal.Width+Petal.Length,
    data=train)
result.lm <- lm(formula=Sepal.Length ~ Sepal.Width+Petal.Length,
    data=test)
result.lm
summary(result.lm)
```

3. 다중 회귀분석

【논문에서 다중 회귀분석 결과 제시 방법】

종속변수	독립변수	표준오차 (Std.Error)	검정통계량(t)	유의확률 (p)
제품만족도	상수	0.130	5.096	6.65e-07 ***
	제품적절성	0.044	15.684	< 2e-16 ***
	제품친밀성	0.039	2.478	0.0138 *
분석 통계량	Multiple R-squared: 0.5975, Adjusted R-squared: 0.5945 F-statistic: 193.8 on 2 and 261 DF, p-value: < 2.2e-16			

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

3. 다중 회귀분석

【논문에서 다중 회귀분석 결과 제시 방법】

- 연구가설1(H1) : '음료수 제품의 적절성은 제품 만족도에 정(正)의 영향을 미친다.'와 연구가설2(H1) : '음료수 제품의 친밀도는 제품 만족도에 정(正)의 영향을 미친다.'를 분석을 위해서 다중 회귀분석을 실시하였다. 분석 결과를 살펴보면 제품 적절성이 제품 만족도에 미치는 영향은 t=15.684, p < 2e-16으로 유의수준 하에서 연구가설1이 채택되었으며, 제품 친밀도가 제품 만족도에 미치는 영향은 t=2.478, p=0.0138로 유의수준하에서 연구가설2가 채택되었다.
- 회귀모형은 상관계수 R=. 0.702으로 독립변수와 종속변수 간에 다소 높은 상관관계를 나타내며, R^2=.594로 독립변수가 종속변수를 59.4% 설명하고 있다. 회귀모형의 적합성은 F=374.020(p-value : < 2.2e-16)으로 나타나서 모형이 적합하다고 볼 수 있다.

3. 다중 회귀분석

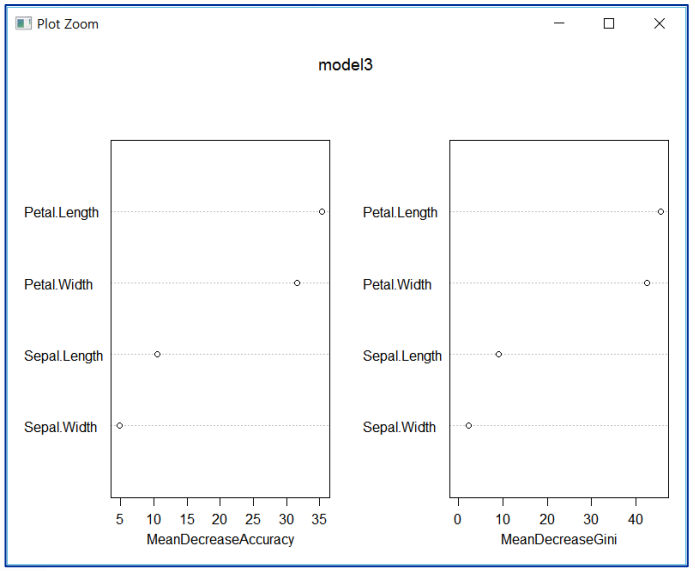
【다중 회귀분석 결과 정리 및 기술】

▪ 가설 설정	연구가설1(H_1) : 음료수 제품의 적절성은 제품 만족도에 **정(正)**의 영향을 미친다.	
	연구가설2(H_1) : 음료수 제품의 친밀도는 제품 만족도에 **정(正)**의 영향을 미친다.	
1. 회귀식 모델 적합성	1) 유의순준	α = 0.05
	2) 검정통계량	*F* = 193.8
	3) 유의확률	*P* = < 2.2e-16
	4) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 회귀선이 모델에 적합하다고 볼 수 있다.
2. 가설검정	1) 유의수준	α = 0.05
	2) 검정통계량	제품적절성(t=15.684), 제품친밀도(t=2.478)
	3) 유의확률	제품적절성(p < 2e-16), 제품친밀도(p=0.014)
	4) 결과해석	유의수준 0.05에서 연구가설이 채택되었다. 따라서 제품 적절성과 제품 친밀도가 높을 수록 제품 만족도가 높아지는 경향을 보이고 있다.

4. 효과적인 변수 선택법

1. 유효하지 않은 x 변수 제거
   - ✓ 선형회귀분석에서 x변수의 유의성 검정 확인
2. 변수 선택법 이용
   - ✓ Stepwise : 전진선택법, 후진제거법, 단계선택법
3. 중요변수 제공 분류 알고리즘
   - ✓ Random Forest : .장버전

```
from xgboost import
XGBClassifier
from xgboost import
plot_importance
```

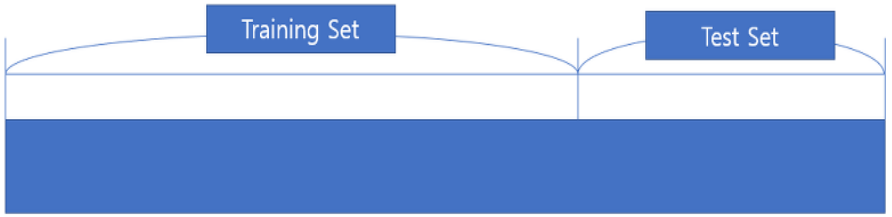


5. 기계학습

1. 홀드아웃 방식

   ✓ 7:3 또는 8:2 비율로 분류한 후 Training set으로 학습

   ✓ Test set으로 model 평가



2. 교차검증

   ✓ n개 균등 분할 후 Train과 Test set을 번갈아 가면서 model을 학습하고 평가하는 방식

	A	B	C	D	E
Cross Validation Iteration 1	Test	Train	Train	Train	Train
Cross Validation Iteration 2	Train	Test	Train	Train	Train
Cross Validation Iteration 3	Train	Train	Test	Train	Train
Cross Validation Iteration 4	Train	Train	Train	Test	Train
Cross Validation Iteration 5	Train	Train	Train	Train	Test

15-2. 로지스틱 회귀분석

Chap15_2_LogisticRegression 수업내용

1. Logit 변환
2. Sigmoid Function
3. 이항 로지스틱 회귀
4. 다항 로지스틱 회귀
5. 오분류표[confusion matrix]

1. Logit 변환

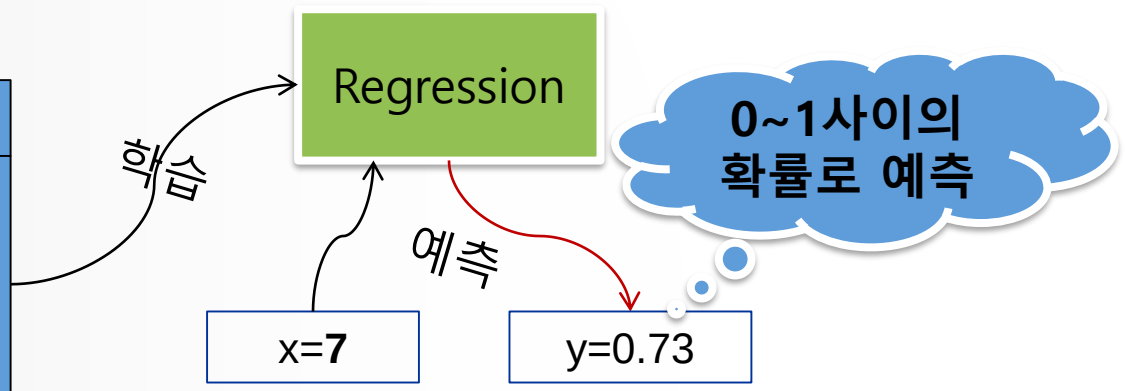
- **오즈비 vs 로짓변환**
  - ## 1. 오즈비(Odds ratio) : 0(실패)에 대한 1(성공)의 비율(0:no, 1:yes)
  - # no인 상태와 비교하여 yes가 얼마나 높은지 or 낮은지 정량화한 것
  - # odds_ratio = p(success) / 1-p(fail)
  - # p : y(반응변수)=1 이 나올 확률, 1-P : y(반응변수)=1의 여사건
  - ## 2. 로짓변환 : 오즈비에 log 함수 적용
  - # logit = log(p / 1-p)

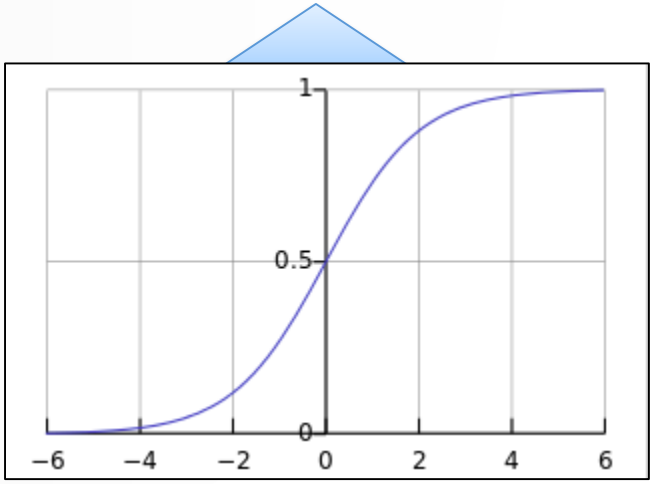
2. Sigmoid Funtion

- **Sigmoid Function**
  - ➢ 합격/불합격 분류

hours	score
10	pass
9	pass
5	fail
3	fail

Training data set





Sigmoid function

3. 이항 로지스틱 회귀

- **이항 로지스틱 회귀모형**

```
#  로지스틱  회귀모델 생성 : 학습데이터
weater_model <- glm(RainTomorrow ~ ., data = train, family = 'binomial')
weater_model
summary(weater_model)

# 로지스틱  회귀모델 예측치 생성 : 검정데이터
# newdata=test : 새로운 데이터 셋, type="response" : 0~1 확률값으로 예측
pred <- predict(weater_model, newdata=test, type="response")
Pred
```

4. 다항 로지스틱 회귀

- **다항 로지스틱 회귀모형**

```
model <- multinom(Species ~ ., data = train)
fit <- model$fitted.values


# type='response' : 0~1 확률 예측 -> sigmoid 함수(yes/no)
# type='probs' : 0~1 확률 예측 -> softmax 함수(a, b, c)
pred_prob <- predict(model, newdata=test, type="probs")
pred_prob
```

5. 오분류표 (confusion matrix)

		예측치 Positive	예측치 Negative
실제값	POS	TP[참 긍정]	FN[거짓 부정]
	NEG	FP[거짓 긍정]	TN[참 부정]

정분류율(Accuracy)= (TP+TN) / 전체관측치

오분류율(Inaccuracy) = (FN+FP) / 전체관측치

정확률(Precision) = TP / (TP + FP)

재현율(Recall) = TP / (TP + FN)

F 측정치(F measure)=2 x $\frac{\text{Precision x Recall}}{\text{Precision + Recall}}$

정분류율(Accuracy) : 알고리즘의 성능평가 척도
오분류율(Inaccuracy) : 알고리즘의 오차 비율
정확률(Precision) : 알고리즘이 Yes로 판단한 것 중에서 실제로 Yes인 비율
재현율(Recall) : 실제값이 Yes인 것 중에서 알고리즘이 Yes로 판단한 비율
F 측정치(F measure) : 정확률과 재현율을 동시에 고려하는 측정치

오분류표(confusion matrix)와 ROC 그래프

예측치

실제값	Positive	Negative
POS	TP[참 긍정]	FN[거짓 부정]
NEG	FP[거짓 긍정]	TN[참 부정]

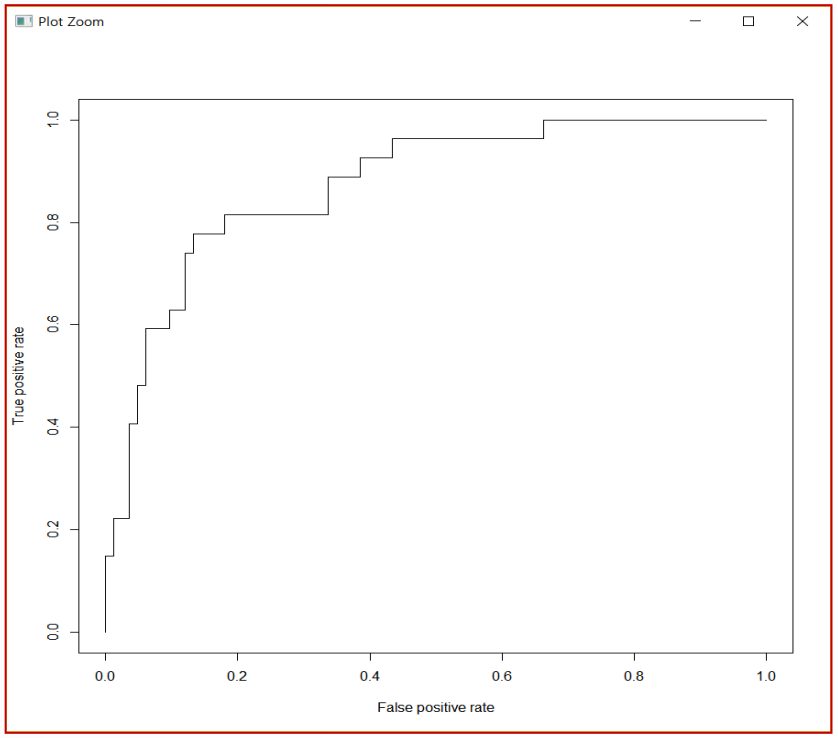
민감도(Sensitivity)= TP / (TP + FN)

특이도(Specificity) = TN / (FP+ TN)

민감도(Sensitivity) : 실제값 Yes인 경우 Yes 예측 비율

= 재현율(Recall)

특이도(Specificity) : 실제값 No인 경우 No 예측 비율



민감도

특이도(Specificity)