데이터 분석 기반
AI 시스템 개발자 양성 과정

05.인공지능 수학 통계 방법론

조창제

강의자료

2025.08.

목차

CONTENTS

01.개요

1.배경

1) 기원

① Statistics는 상태를 뜻하는 Status라는 고대 라틴어에서 출발하였으며, 정치관련 용어로 시작됨

② 게임에서 Status는 나의 강한 정도를 나타내는 **정량적 수치**를 의미

③ 최초의 통계 추론

- 18~19세기 출생 신고 및 사망자 수를 통한 인구 추론에 사용

2) 정의

① 관심을 갖는 어떤 대상에 대하여, 이를 정리/요약하고 자료를 분석하여 불확실한 사실에 대하여 과학적 합리적 판단을 내리는 방법

3) 구분

① 기술통계학(Descriptive Statistics): 자료를 **특정 수치로 정리/요약**에 사용하는 분야

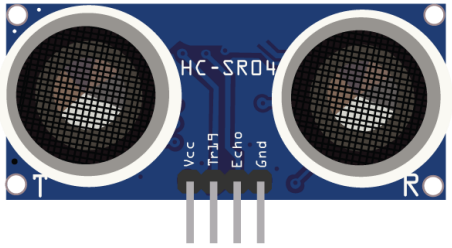
② 추론통계학(Inferential Statistics): 자료를 분석하여 **통계적인 판단**(가설 검증)에 사용하는 분야

2.정의

1) 측정

① 측정(Measurement)

- **어떤 대상이나 현상의 속성**을 수치로 나타내기 위해 **특정한 기준이나 도구**를 사용하여 **수치를 부여하는 절차**
- 타당도(Validity): 검사 도구가 **측정하려는 특성을 충실히 측정**하고 있는지의 정도
- 신뢰도(Reliability): 측정하고자 하는 것을 **일관되고 정확**하게 측정하는 있는지의 정도



센서

저울

시계

온도계

2.정의

1) 측정

② 척도(Scale)

- 어떠한 대상이나 현상의 속성을 수량적으로 측정하거나 평가하기 위한 기준

척도명	정의	예시	순서 (Order)	크기 (Magnitude)	비율 (Ratio)	연산
명목척도 (Nominal scale)	단순히 이름이나 범주만 구분 측정 대상을 구분하기 위해 수치를 부여한 것	성별, 국적, 혈액형	X	X	X	독립성 검정, 적합도 검정, 로지스틱회귀
서열척도 (Ordinal scale)	서열이 있는 측정대상을 수치로 구분한 것 간격이 동일하지 않음	학년, 설문조사 만족도, 호텔등급, 영화평점	O	X	X	스피어만 상관
등간척도 (Interval scale)	서열이 있는 측정대상을 동일한 간격으로 구분한 것 간격이 일정하지만 절대적 0이 없음	온도, IQ 점수, 연도	O	O	X	변동계수(X)
비율척도 (Ratio scale)	0이 개념이 존재하는 등간척도	무게, 키, 수익, 나이	O	O	O	모든 통계량 가능

2.정의

2) 모집단(Population)

① 연구 또는 조사의 대상이 되는 전체 집단을 의미

② 모수(Parameter)

- 모집단의 특성을 나타내는 수치

3) 표본(Sample)

① 모집단에서 일부를 선택하여 이루어진 부분집합

구분	정의
표본공간(Sample space)	실험이나 관찰에서 가능한 모든 결과들의 집합(범주)
사건(Event)	표본 공간 내에서 발생할 수 있는 특정 결과들의 부분집합
통계량(Statistic)	모집단의 모수를 추정하기 위한 값(표본에서 계산된 수치)

2.정의

3) 예시

① 한 대학교 입학처에서 신입생들의 수학 능력을 평가하기 위해 수학시험을 실시

- 모집단: 이 대학교에 입학한 5,000명 신입생 전체
- 표본: 무작위로 선택된 500명의 신입생
- 표본공간: 시험점수의 가능한 모든 값(0~100점)
- 사건
  - 시험점수가 90점 이상인 경우
  - 시험점수가 60점 미만인 경우
  - 시험점수가 70~80점 사이인 경우
- 통계량: 표본의 평균점수(75.3점)

2.정의

4) 표본추출(Sampling)

① 전체 모집단에서 일부를 선택하여 데이터를 수집하고, 모집단에 대한 추론이나 분석을 수행하는 과정

② 비용절감, 시간절약을 목표로 함

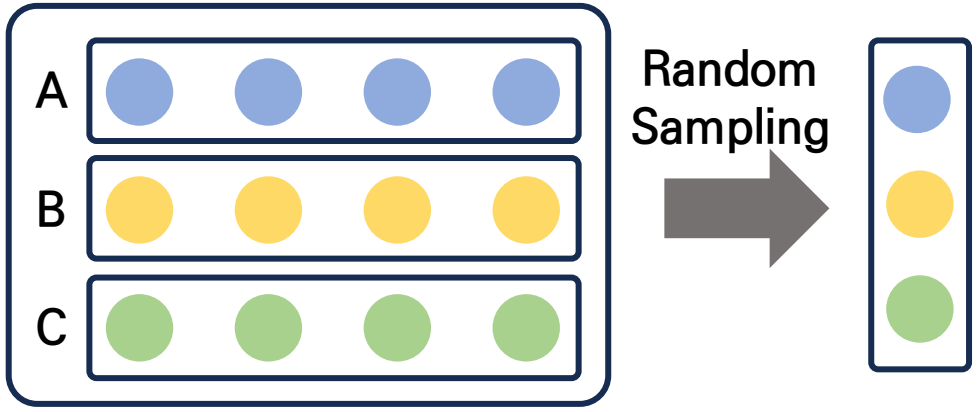
척도명	정의	대표성	효율성
단순추출 (Simple random sampling)	모집단에서 표본을 무작위로 선택하는 방법	높음	낮음
계통추출 (Systematic sampling)	모집단을 **일정한 간격**으로 나눈 후, 첫 표본을 무작위로 선택하고, 그 후 일정 간격마다 표본을 추출하는 방법	보통	높음
층화추출 (Stratified sampling)	모집단을 **서로 유사한 특성을 가진 그룹**으로 나눈 후, 각 그룹마다 무작위 표본을 추출하는 방법	높음	낮음
집락추출 (Cluster sampling)	모집단을 이질적인 여러 개의 군집으로 나눈 후, **선택된 특정 군집의 모든 표본** 혹은 **일부**를 제시하는 방법	낮음	높음

2.정의

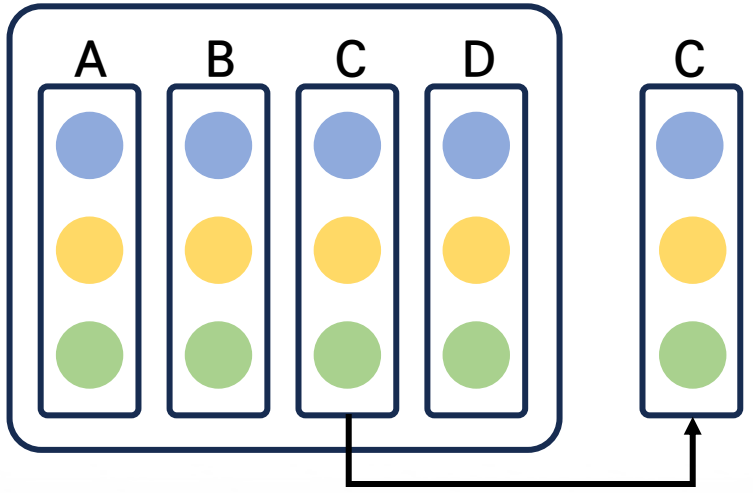
4) 표본추출(Sampling)

① 전체 모집단에서 일부를 선택하여 데이터를 수집하고, 모집단에 대한 추론이나 분석을 수행하는 과정

② 비용절감, 시간절약을 목표로 함



층화 추출



집락 추출

2.정의

5) 변수(Variable)

① 개체나 관측 단위마다 다른 값을 가질 수 있는 특성이나 속성

6) 상수(Constant)

① 모든 관측 단위에 대해 동일한 값을 가지는 특성이나 속성

인과관계에 따른 변수의 구분	정의
독립변수(Independent)	**연구자가 조작**하거나 선택하는 변수
종속변수(Dependent)	독립변수의 **영향을 받아 변화**하는 변수

속성에 따른 변수의 구분	정의	예시
질적변수(Qualitative)	범주나 속성을 나타내는 변수(수량화 불가)	성별, 혈액형
양적변수(Quantitative)	수량으로 표시되는 변수	키, 몸무게, 소득
이산형 변수(Discrete)	셀 수 있는 값을 가지는 변수	자녀 수
연속형 변수(Continuous)	특정 범위 내에서 무한히 많은 값을 가질 수 있는 변수	시간, 온도

2.정의

7) 확률(Probability)

① 어떤 사건(Event)이 발생할 가능성을 나타낸 값

② 확률변수(Random variable)

- 표본 공간의 각 결과에 실수 값을 대응시키는 함수
- 예시 – 주사위 눈, 동전 앞/뒤, 몸무게, 온도
- 이산확률변수(Discrete random variable)
- 연속확률변수(Continuous random variable)

구분	정의
이산확률변수 (Discrete random variable)	유한하거나 셀 수 있는 무한한 개수의 값만 취하는 확률변수
연속확률변수 (Continuous random variable)	특정 구간 내의 모든 실수 값을 취할 수 있는 확률변수

2.정의

7) 예시

① 주사위 1개를 던지는 실험

- 표본공간: 1, 2, 3, 4, 5, 6
- 확률변수: 주사위 눈의 수
  - X(1) = 1, X(2) = 2, ..., X(6) = 6
- 확률변수: 짝수이면 1 홀수이면 0
  - Y(1) = 0, Y(2) = 1, ..., Y(6) = 1

② 동전던지기

- 확률변수: 앞면(H)이 나온 횟수
  - X(HHH)=3, X(HHT) =2 ,..., X(TTT) = 0

2.정의

7) 확률(Probability)

③ 조건부확률(Conditional probability)

- 사건B가 일어났을 때 사건 A가 일어날 확률
- $P(A|B) = {}^{P(A\cap B)}/_{P(B)}$

구분	정의
독립(Independent)	어떤 한 사건이 일어나든 일어나지 않든, 다른 사건에 영향을 주지 않으면 독립 (즉, A와 무관하게 B의 확률이 같은 경우) $P(B\|A) = {}^{P(A\cap B)}/_{P(A)} = {}^{P(A)P(B)}/_{P(A)} = P(B)$
종속(Dependent)	사건 A의 발생이 사건 B의 확률에 영향을 미치는 경우

2.정의

7) 확률(Probability)

③ 조건부확률(Conditional probability)

- 사건B가 일어났을 때 사건 A가 일어날 확률
- $P(A|B) = {}^{P(A\cap B)}/_{P(B)}$
- θ 값이 주어지고, 사건이 발생할 확률을 계산하는 것을 의미

- 사건 A: 10번 동전 던져서 앞면이 7번 나옴
- 사건 B: 동전이 앞면일 확률($\theta = 0.5$)
- θ=0.5 라 할 때, 앞면이 7번 나올 확률은?

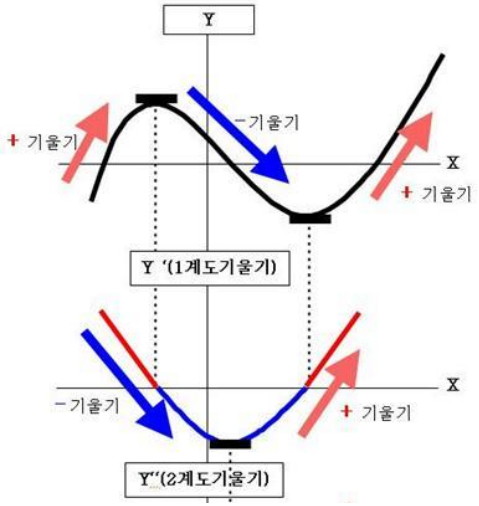
$$P(X = 7|\theta) = \binom{10}{7}\theta^7(1-\theta)^3 = \frac{10!}{7!\,3!}\theta^7(1-\theta)^3 = 120\,\theta^7(1-\theta)^3$$

2.정의

7) 확률(Probability)

③ 우도함수(Likelihood Function)

- 관측된 데이터를 고정하였을 때, 모수의 값을 변수로 하는 함수를 의미
- $L(\theta|X) = P(X|\theta)$
- 동전 10번 던졌을 때 앞면이 7번 나왔을 때, 가장 잘 나타날 법한 확률(불확실성 최소화)
  - $L(\theta)$가 최대가 되게 하는 θ값을 찾는 것



$$L(\theta|X) = P(X=7|\theta) = \binom{10}{7}\theta^7(1-\theta)^3 = \frac{10!}{7!\,3!}\theta^7(1-\theta)^3$$

$$l(\theta) = logL(\theta) = \log 120 + 7\log\theta + 3\log(1-\theta)$$

$$\frac{d}{d\theta}l(\theta) = \frac{7}{\theta} - \frac{3}{1-\theta} = 0$$

$$7(1-\theta) - 3\theta = 0$$

$$7 - 10\theta = 0$$

$$\theta = 0.7$$

우도함수는 확률이므로 항상 양수
로그는 계산 편의성 때문에 사용

x가 0보다 크면 로그는 항상 단조 증가 함수

$$x_1 < x_2\,, \log x_1 < \log x_2$$

$$l''(\theta) = -\frac{7}{\theta^2} - \frac{3}{(1-\theta)^2} < 0$$

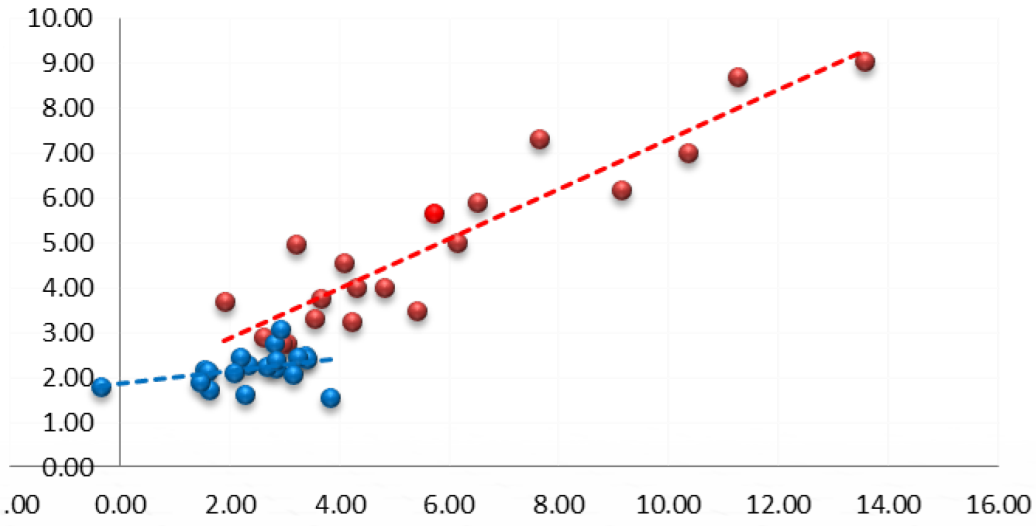
2계 도함수 값이 음수이므로 최대값

1.정의

1) 중심 경향 척도

① 최빈값(Mode): 데이터 집합 중 가장 자주 나타나는 값
② 평균(Mean): 데이터 값의 합을 데이터 개수로 나눈 값
③ 중앙값(Median): 데이터를 크기 순으로 정렬했을 때 중앙에 위치한 값

$$mean(X) = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



1.정의

2) 산포(변동성) 경향 척도

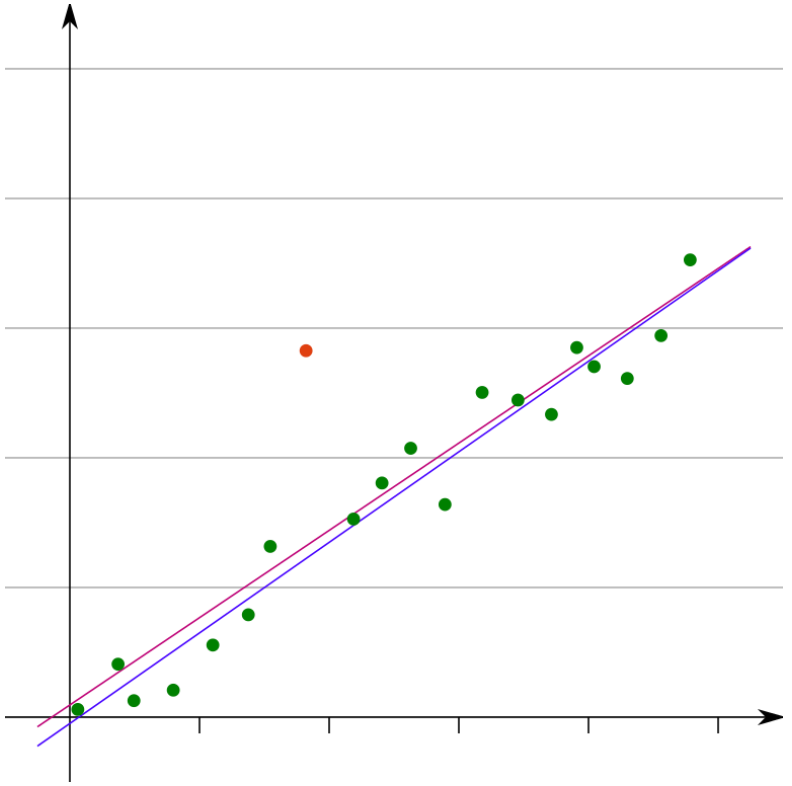
① 범위(Range)
- 데이터 집합에서 최댓값과 최소값의 차이

② 분산(variance, Var)
- 데이터들이 평균으로부터 얼마나 차이가 나는지 나타낸 정도

③ 표준편차(Standard deviation, SD)
- 분산의 제곱근 값으로, 데이터와 같은 단위를 가짐

④ 사분위수 범위(Interquartile range, IQR)
- Q3-Q1

⑤ 변동계수(Coefficient of variation, CV)
- 표준편차를 평균으로 나눈 값



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

1.정의

1) 추론통계

① 표본을 분석하여 모집단의 특성을 추정

- 모집단 전체를 분석하는 것이 불가능하거나 실용적이지 못할 때 사용

2) 가설검정(Hypothesis)

① 어떠한 사실을 설명하거나 어떤 이론 체계를 연역하기 위하여 설정한 가정

구분	정의
귀무가설 (Null hypothesis)	연구에서 검증하려는 기본 가설
대립가설 (Alternative hypothesis)	귀무가설과 반대되는 주장
유의수준 (Significance level)	귀무가설을 기각할 때, 잘못 기각하는 오류의 확률(1종오류의 확률)

1.정의

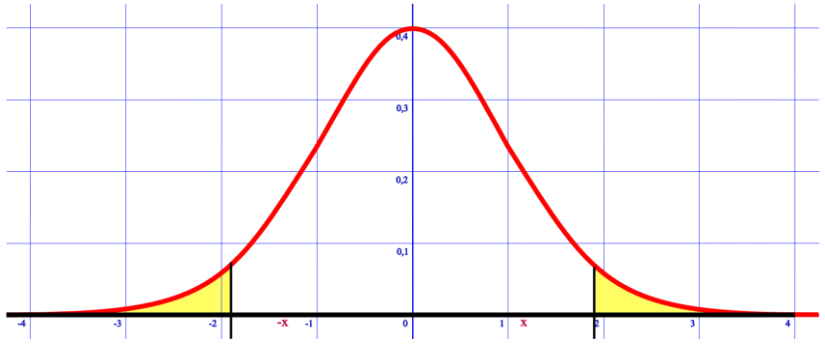
3) 가설검정(Hypothesis)

① 어떠한 사실을 설명하거나 어떤 이론 체계를 연역하기 위하여 설정한 가정

② 가설검정의 방식

- 유의수준 설정 〉 통계량 계산 〉 p값

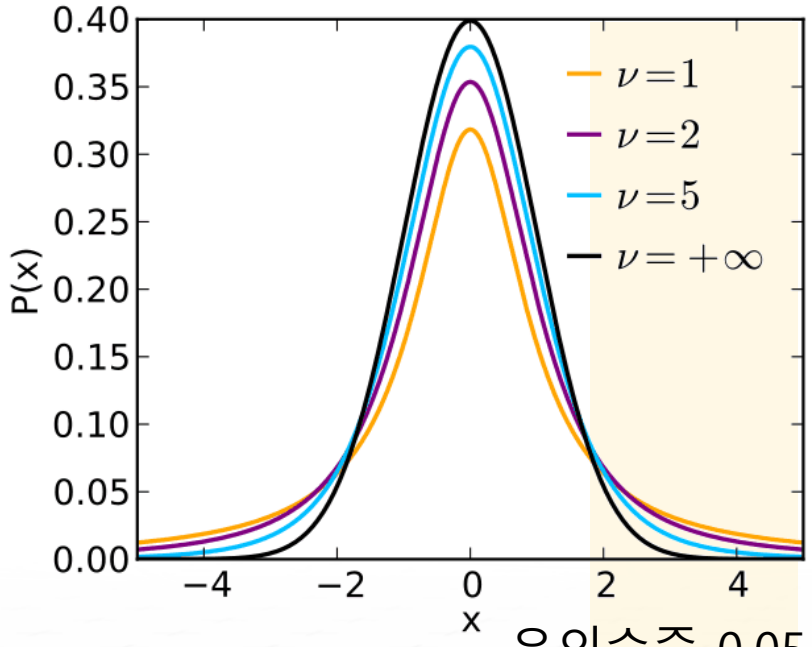
구분	정의
양측검정	검정 통계량이 두 방향에서 극단적인 값일 때 귀무가설을 기각
단측검정	검정 통계량이 한 방향에서만 극단적인 값일 때 귀무가설을 기각
1종오류(α)	귀무가설이 참인데 귀무가설을 기각하는 오류, 유의수준
2종오류(β)	대립가설이 참인데 귀무가설을 기각하지 않는 오류
p value	귀무가설이 참일 때, 검정 통계량보다 더 극단적인 값이 관찰될 확률



1.정의

3) 가설검정(Hypothesis)

① 어떠한 사실을 설명하거나 어떤 이론 체계를 연역하기 위하여 설정한 가정



유의수준 0.05 기준
(기각역)

T 통계량 기준

H0: 모집단의 평균이 50이다.
H1: 모집단의 평균이 50이 아니다.

$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{\bar{x}-50}{s/\sqrt{n}}$

데이터 100건이라 가정
자유도(df) = 99
표본평균$(\bar{x})$ = 52
표본표준편차(s) = 5
t 통계량 = (52-50)*2 = 4

t(0.05, 99) ≈ 1.660
t(0.025, 99) ≈ 1.984

유의수준 0.05 하에서
귀무가설을 기각
모집단의 평균이 50이라
할 수 없다.

1.정의

4) 모수 통계

① 점추정: 표본평균, 표본분산, 표본비율
② 구간추정: 신뢰구간
③ 가설검정 예시

가설검정	귀무가설	용도	가정
z-test	$\mu = \mu_0$	모집단의 표준편차를 알고 있을 때, 평균의 차이 검증에 사용	모집단 정규분포 가정, 표본 간의 독립성 가정
One sample t-test	$\mu = \mu_0$	표본의 평균이 특정값과 통계적으로 차이가 있는지 검증할 때 사용	
Two sample t-test (Independent)	$\mu_1 = \mu_2$	별개의 두 집단을 비교 검증할 때 사용 (남학생과 여학생의 성적 비교)	
Paired t-test	$\mu_A - \mu_B = 0$	같은 집단의 전후 측정치 비교 검증에 사용 (사전-사후 값의 차이에 대한 일 표본 t 검정)	
Chi-square test	집단의 분포가 동일	두 범주형 변수 간에 분포가 동일한지 검증	범주의 빈도가 최소 5이상
F test	${\sigma_1}^2 = {\sigma_2}^2$	분산의 차이 검증에 사용	모집단 정규분포 가정
ANOVA	$\mu_1 = \mu_2 = \cdots = \mu_k$	그룹 간 평균의 차이가 있는지 검증하는데 사용	모집단 정규분포 가정, 등분산성 가정

1.정의

5) 비모수 통계

① 모집단의 분포적 특성을 가정하지 않음

② 가설검정 예시

가설검정	귀무가설	용도
Sign Test	중앙값이 동일	One sample t test에 대응되는 비모수적 방법
Mann Whitney U Test	중앙값이 동일	Two sample t test에 대응되는 비모수적 방법
Wilcoxon Signed Rank Test	중앙값 차이가 0	Paired t test에 대응되는 비모수적 방법
Kruskal Wallis Test	모든 그룹의 분포가 동일	ANOVA의 비모수적 대체 방법

1.정의

1) 선형회귀

① 종속변수를 가장 잘 맞추는 선을 찾는 것을 의미

② 최적화 기준

- 최소제곱법: 오차가 최소가 되는 직선을 찾는 방법
- 최대우도법: 가장 일어날 법한 직선을 찾는 방법

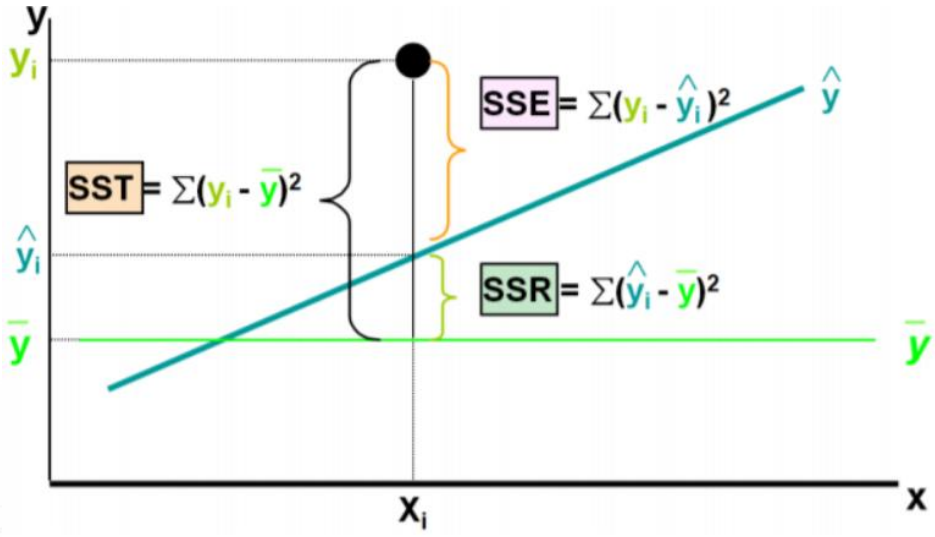
가정	설명	검정
선형성	독립변수와 종속변수 사이에는 선형 관계	산점도, 잔차그림, Ramsey Reset Test
등분산성	오차의 분산이 동일 (예측값이 크거나 작음에 관계없이 오차의 크기가 일정)	Bartlett's Test, Levene's Test
독립성	잔차 독립 (시간의 변화에 관계없이 오차의 크기가 일정, 주기성이 없음)	Durbin Watson Test
정규성	오차는 자연스러운 분포(정규분포)를 따른다 가정	Shapiro-wilk Test, Anderson-darling Test, Kolmogorov-Smirnov Test(KS Test)
(비다중공선성)	독립 변수 간에 상관관계가 없음	분산팽창지수(Variance Inflation Factor, VIF)

1.정의

1) 선형회귀

③ 모형해석

- SST: 종속변수 y의 총 변동, 평균 대비 얼마나 움직였는지에 대한 정도, 종속변수 Y의 총 변동량
- SSR: 독립변수에 의해 예측 값이 평균에서 얼마나 움직였는지에 대한 정도, 모형이 설명한 변동
- SSE: 예측 값과 실제 값의 차이, 모형이 설명하지 못한 변동
- 결정계수 = SSR/SST = 1-SSE/SST
  - 전체 변동량 중 모형이 설명한 변동량

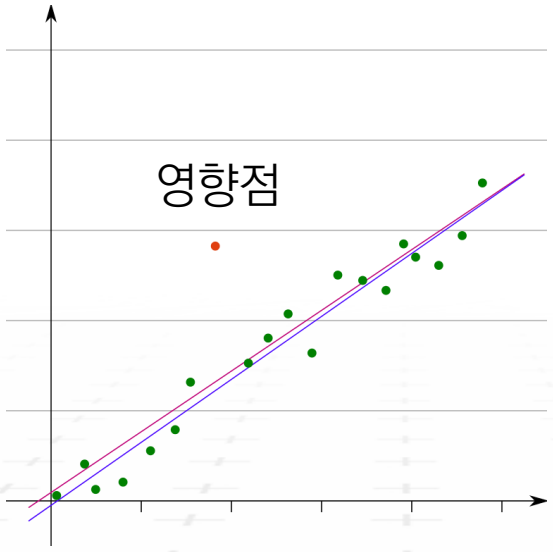


1.정의

1) 선형회귀

③ 모형진단

- 지레점(Leverage points): 독립변수 값이 평균에서 멀리 떨어져 있는 데이터
  - 영향점(Influential points): 회귀 모형의 추정 결과에 큰 영향을 미치는 데이터
- 쿡의거리(Cook's Distance)
  - 각 데이터점이 회귀분석의 결과에 미치는 영향을 측정하는 지표
- F검정 – H0: 모든 회귀계수는 모두 0이다.
- T검정 – H0: 회귀계수는 0이다.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.416
Model:	OLS	Adj. R-squared:	0.353
Method:	Least Squares	F-statistic:	6.646
Date:	Sun, 18 Feb 2018	Prob (F-statistic):	0.00157
Time:	15:18:50	Log-Likelihood:	-12.978
No. Observations:	32	AIC:	33.96
Df Residuals:	28	BIC:	39.82
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.4639	0.162	2.864	0.008	0.132	0.796
x2	0.0105	0.019	0.539	0.594	-0.029	0.050
x3	0.3786	0.139	2.720	0.011	0.093	0.664
const	-1.4980	0.524	-2.859	0.008	-2.571	-0.425

Omnibus:	0.176	Durbin-Watson:	2.346
Prob(Omnibus):	0.916	Jarque-Bera (JB):	0.167
Skew:	0.141	Prob(JB):	0.920
Kurtosis:	2.786	Cond. No.	176.

1.정의

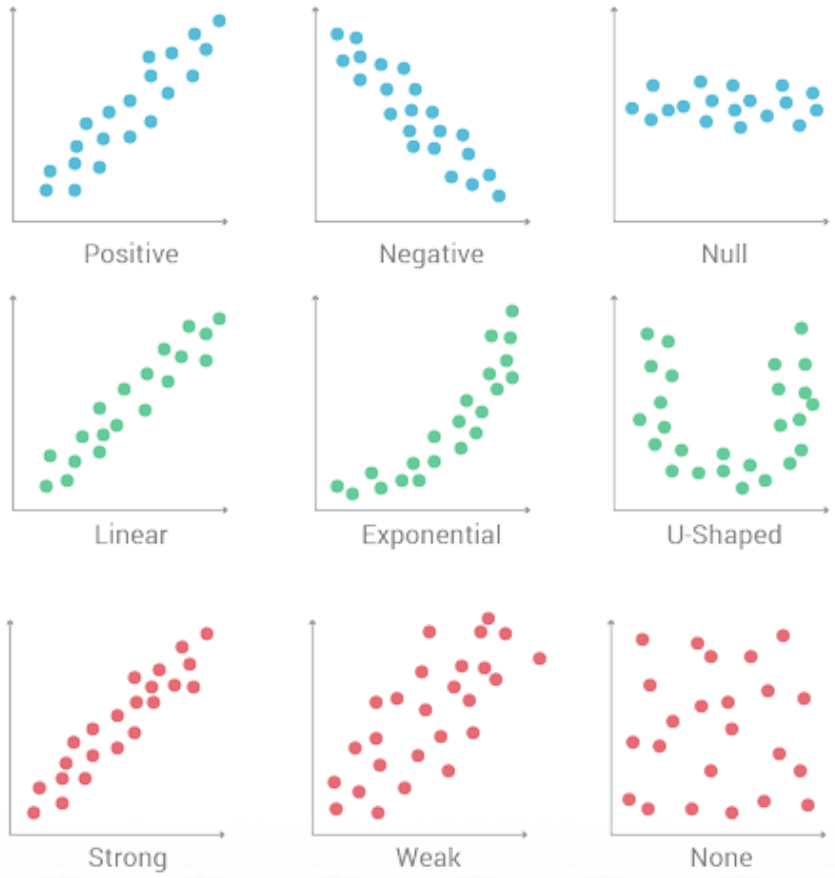
2) 상관분석

① Pearson correlation coefficient
- 모수적 방법
- 정규 분포 가정
- 두 변수 사이의 선형적인 상관성이 있는지 검증

② Spearman correlation coefficient
- 비모수적 방법
- 두 변수 사이의 순위가 상관성이 있는지 검증

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)}}$$



상관성 예시

1.정의

3) 회귀모형 구분

① 일반화 선형 모형(Generalized linear model)

- 선형 모델을 확률 분포를 기반으로 일반화한 모델을 의미
- $g(\mu) = X\beta$
- 종속변수가 확률(0~1범위), 개수(0 혹은 양의 정수), 시간의 변화에 따라 특정 패턴이 존재하는 경우에 사용
- 연결함수: 종속변수의 적용하는 함수. $g(\mu)$에 해당

가정	설명	연결함수
로지스틱 회귀	이진 분류 문제에 활용	$\log {}^{\mu}/_{1-\mu}$
포아송 회귀	개수 데이터 예측에 활용 (방문자 수, 사고 건수)	$\log(\mu)$
감마 회귀	양수 연속형 데이터 예측에 활용 (보험료, 대기시간)	$\log(\mu)$, ${}^{1}/_{\mu}$
지수 회귀 (생존분석)	시간 데이터를 예측 (제품의 수명, 고객 이탈 시간)	$\log(\mu)$

1.정의

3) 회귀모형 구분

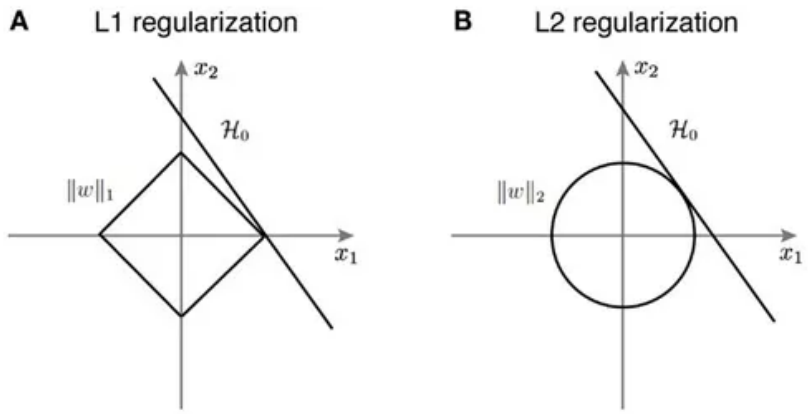
② 일반화 회귀 모형(Generalized regression model)

- Lasso Regression
  - L1 Regularization: 손실함수에 회귀계수의 절대값의 합을 추가
  - 회귀 계수 중 일부를 0으로 만든다는 특징 존재
- Ridge Regression
  - L2 Regularization: 손실함수에 회귀계수의 제곱합을 추가
  - 회귀 계수 중 일부를 0에 가깝게 만들지만 완전히 0으로 만들지는 않음
- Elastic net(L1과 L2의 혼합)

$$L(\beta) = \sum_{i=1}^{n}(y_i - X_i\beta)^2$$

$$L(\beta) = \sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$L(\beta) = \sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$



1.정의

4) 모형비교

① 설명력이 좋고, 좋은 모델의 기준은 불필요한 변수가 적은 모델

② 설명력(Explained Variance)

- 독립변수가 종속변수를 얼마나 잘 설명하는지를 나타내는 지표
- 총 변동(SST) 중에서 모델이 설명한 변동(SSR = SST-SSE)

③ 우도함수(Likelihood): 주어진 데이터가 특정 모델에서 나올 확률의 의미로 해석

지수	설명	수식
결정계수	모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{SSR}{SST}$
수정결정계수	변수 개수에 따른 보정을 적용한 결정계수 (p는 변수 개수)	$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-p-1}\right]$
AIC (Akaike information criterion)	모델 적합성과 복잡도를 균형있게 평가 변수를 최소로 쓰고 불확실성을 최소화	$AIC = 2p - 2\ln(L)$, L은 우도함수
BIC (Bayesian information criterion)	AIC보다 모델 복잡도에 강한 패널티를 적용	$AIC = pln(n) - 2\ln(L)$, L은 우도함수

1.정의

5) 변수선택법(Feature selection)

가정	설명
전진선택법 (Forward selection)	1. 모든 변수에 대해 해당변수 하나만 넣고 회귀분석을 수행 2. p value가 가장 낮은 변수를 고정적으로 추가 3. **나머지 변수를 하나씩 넣어보며** 회귀분석을 수행 4. 변수 추가 전/후 비교하여 모델 성능이 개선되면 변수 추가 5. 성능이 개선되지 않으면 최종 모델 제시
후진제거법 (Backward elimination)	1.모든 변수를 모두 넣고 회귀분석을 수행 2.모든 변수에 대해 **해당변수 하나를 빼고** 회귀분석을 수행 3.변수 제거 전/후 비교하여 모델 성능이 개선 시 변수제거 4.성능이 개선되지 않으면 최종 모델 제시
단계적방법 (Stepwise selection)	1.전진선택법처럼 첫번째 변수 선택 2.새로운 변수를 **넣는 경우와 변수를 빼는 경우 모두의 경우에 대해** 회귀분석을 수행 3.모델 성능이 개선되지 않으면 최종 모델 제시
Boruta (비 통계적방법)	1. 기존변수와 기존변수를 **복원추출해 만든 변수(그림자변수)**들로 Random Forest 모델 생성 2. 생성된 모형의 변수중요도에서 **그림자변수 중 가장 변수 중요도가 높은 값**보다 변수중요도가 **높게 나온 기존변수**에 대해 **HIT 수를 1씩 증가** 3. HIT 수를 기준으로 순위 검정을 통해 기각 유무를 분류 4. 모든 변수가 분류되면 최종 변수를 제시

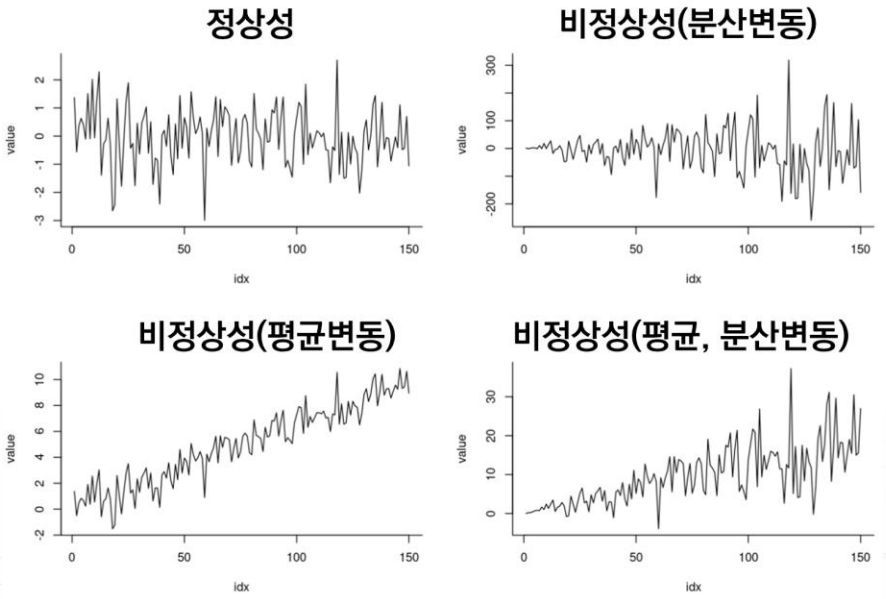
1.정의

1) 시계열 자료의 정의

① 주기성이 뚜렷한 자료로 자기상관, 트랜드, 계절성, 주기성 등을 포함하는 자료

② 정상성(Stationarity)

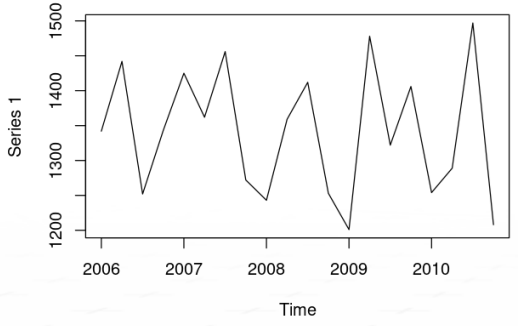
- 시계열 데이터의 통계적 특성이 시간에 따라 변하지 않는 성질을 의미
- 시간이 지나도 평균이 일정하고 분산이 일정
- 더빈 왓슨 통계량 값이 2에 가까운지로 잔차의 자기상관이 없는지 판정
  - 잔차의 자기상관 존재: 특정 시차에 따른 의존성이 존재



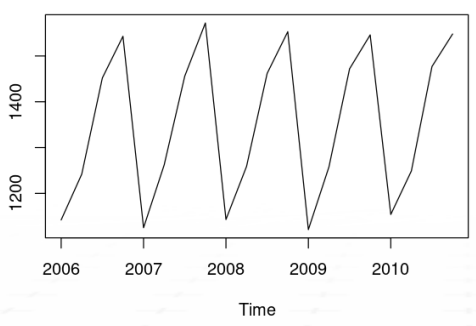
1.정의

2) 시계열 자료의 구분

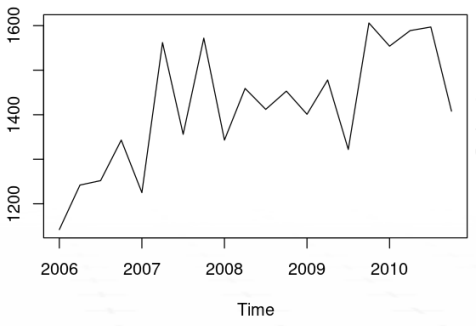
구분	설명	수식
우연변동(random)	시계열 자료가 일정 수준에 머물며 우연변동에 의한 변화만 나타내는 경우	$y_t = \alpha + e_t$
계절변동(seasonal)	시계열 자료가 주기적 성격의 계절변동에 의한 영향을 받아 주기적인 변화를 나타내는 경우 주기가 긴 계절변동은 순환변동(Cyclical variation)이라 함	$y_t = \alpha + \beta_1 sin\frac{2\pi t}{f} + \beta_2 cos 2\pi t$ $f + e_t$
추세변동(trend)	시간에 따라 증가하거나 감소하는 경향을 보여주는 경우	$y_t = \alpha + \beta t + e_t$ $y_t = \alpha + \beta_1 t + \beta_2 t + e_t$
계절적추세변동 (seasonal trend)	추세변동과 계절변동을 동시에 보여주는 경우	$y_t = \alpha + \beta_1 t + \beta_2 sin\frac{2\pi t}{f} + cos\frac{2\pi t}{f} + e_t$



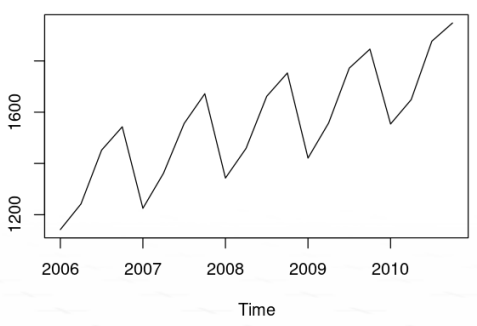
우연변동



계절변동



추세변동



계절적추세변동

1.정의

3) BoxCox변환

① 긴 꼬리 형태의 비대칭 분포를 정규분포에 가깝게 변환

② 오차의 분산이 안정화되는 효과를 가짐

$$f(x:\lambda) = \frac{x^{\lambda}-1}{\lambda}, (\lambda \neq 0)$$

$$f(x:0) = ln(x), (\lambda = 0)$$

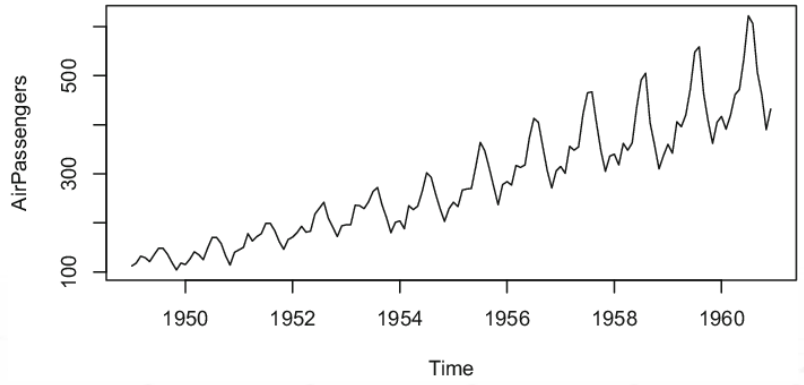
BoxCox 변환 수식

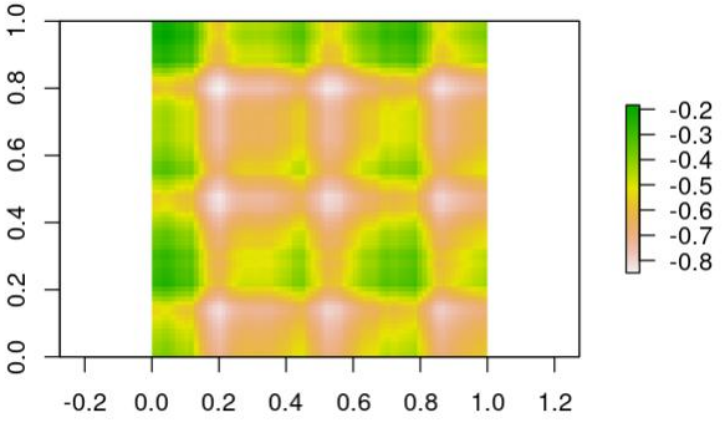
4) 모형

① ARIMA(Auto Regressive Integrated Moving Average)

- 차분으로 정상성을 고려, 과거의 값과 과거의 오차 모두 현재 값을 정하는데 기여한다고 가정하는 모형

② GARCH, VAR, ETS 등 다양한 모형이 존재





그래프를 이미지화 해서 분석하기도 함

1.정의

4) 자기상관(Auto correlation)

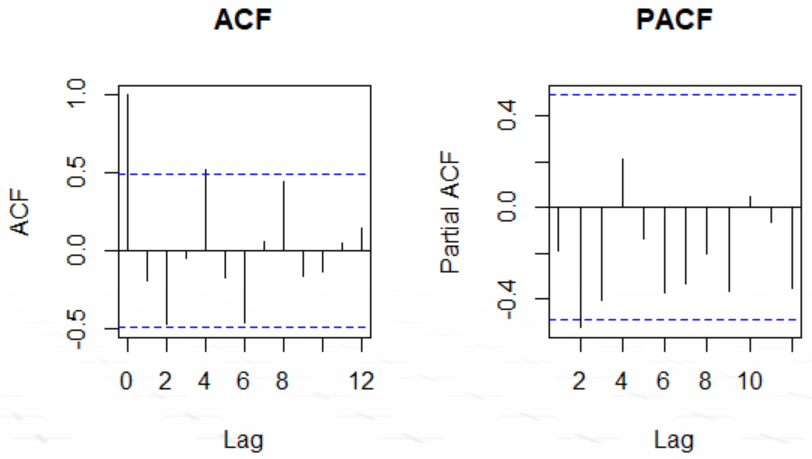
① 현재시점과 k 시점 전의 관측값 사이의 상관계수

5) 부분자기상관(Patrial auto correlation)

① 다른 중간시점들의 영향을 제외한, 두 시점 사이의 순수한 상관성을 의미

② 순수한 상관성

- 중간 시점의 값을 독립변수로 현시점을 종속변수로 한 잔차를 계산
- 중간 시점의 값을 독립변수로 k시점 전의 관측값을 종속변수로 한 잔차를 계산
- 부분자기 상관은 잔차들 사이의 상관계수를 의미



1.정의

1) 스칼라(Scalar)

① 크기만 존재하는 값

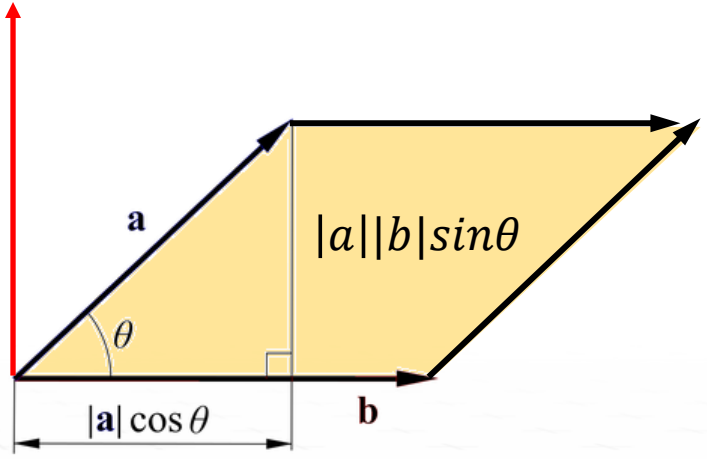
2) 벡터(Vector)

① 크기와 방향을 가지는 값

3) 내적

① 같은 방향이면 양수, 반대방향이면 음수, 수직이면 0의 값을 가짐

구분	설명	수식
내적 (Dot product)	두 벡터가 같은 방향으로 얼마나 정렬 되어 있는지를 의미 (즉, 두 벡터 사이의 사영을 의미)	$a \cdot b = \|a\|\|b\|cos\theta$
외적 (Cross product)	두 벡터가 이루는 평면과 수직인 단위벡터를 의미 (평행사변형의 넓이)	$a \times b = \|a\|\|b\|sin\theta$



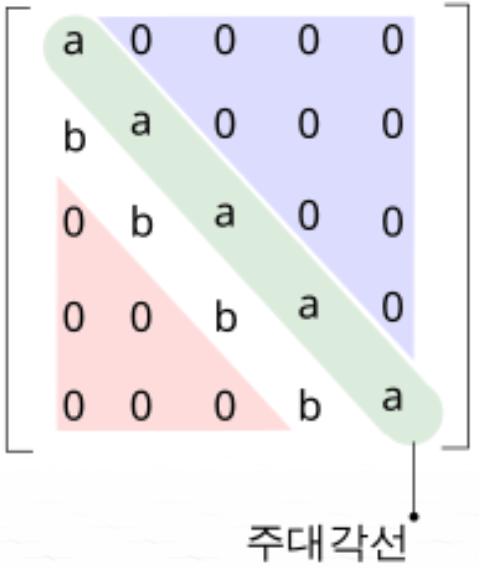
내적과 외적

1.정의

1) 행렬(Matrix)

① 수 또는 문자를 직사각형으로 배열하고 괄호로 묶어 나타낸 것
② 행(Row): 가로 줄의 성분 전체
③ 열(Column): 세로 줄의 성분 전체
④ 성분(Element, Component): 괄호 내의 수 또는 문자

구분	설명
주대각성분 (Main diagonal element)	n차 정사각행렬에서 대각선의 성분
상삼각행렬 (Upper Triangular)	n차 정사각행렬에서 i>j 에 대하여 0값을 가지는 행렬
하삼각행렬 (Lower Triangular)	n차 정사각행렬에서 i<j 에 대하여 0값을 가지는 행렬
대각행렬(Diagonal)	주 대각성분 이외의 값이 모두 0인 행렬
전치행렬(Transposed)	행과 열을 바꾸어 놓은 행렬



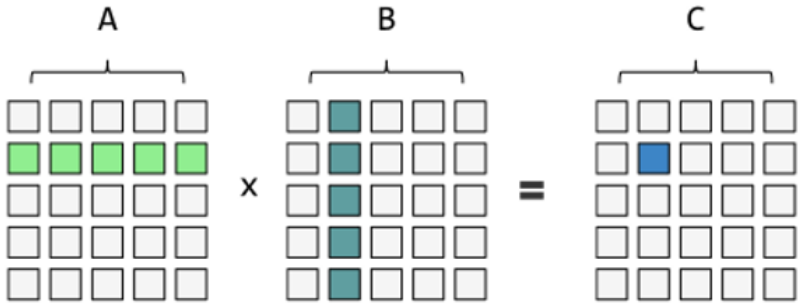
1.정의

1) 행렬(Matrix)

⑤ 행렬의 곱셈

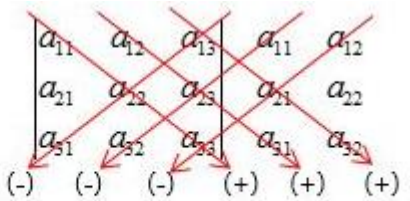
⑥ 행렬식(Determinant)

- 행렬의 공간의 크기 변화량을 의미함
- 2차원에서는 외적과 동일한 개념



행렬곱

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

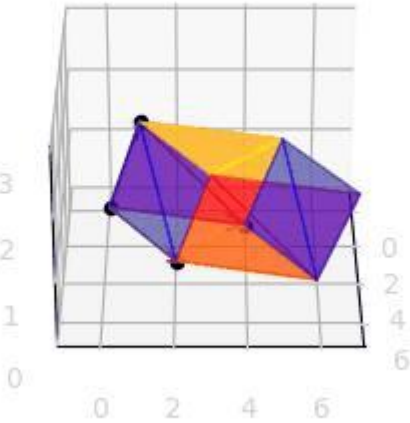


$$= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

행렬식

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \quad \begin{matrix} (a & b & c) \\ (d & e & f) \\ (g & h & i) \end{matrix}$$

3개의 벡터



행렬식의 가시화

1.활용

1) 아핀변환(Affine Transformation)

① 2D나 3D 공간에서 점, 선, 도형을 변환 및 이동하는 방법 중 하나

② 이미지의 arr[x, y]의 값을 다시 new_arr[new_x, new_y]에 할당

③ x, y 와 변환 행렬 사이의 행렬곱 연산 수행

- 이동(Translation)
- 확대/축소(Scaling)
- 회전(Rotation)
- 반사(Reflection)
- 비틀기(Shearing)

$$T = \begin{pmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}$$

이동

$$S = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix}$$

확대/축소

$$R = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

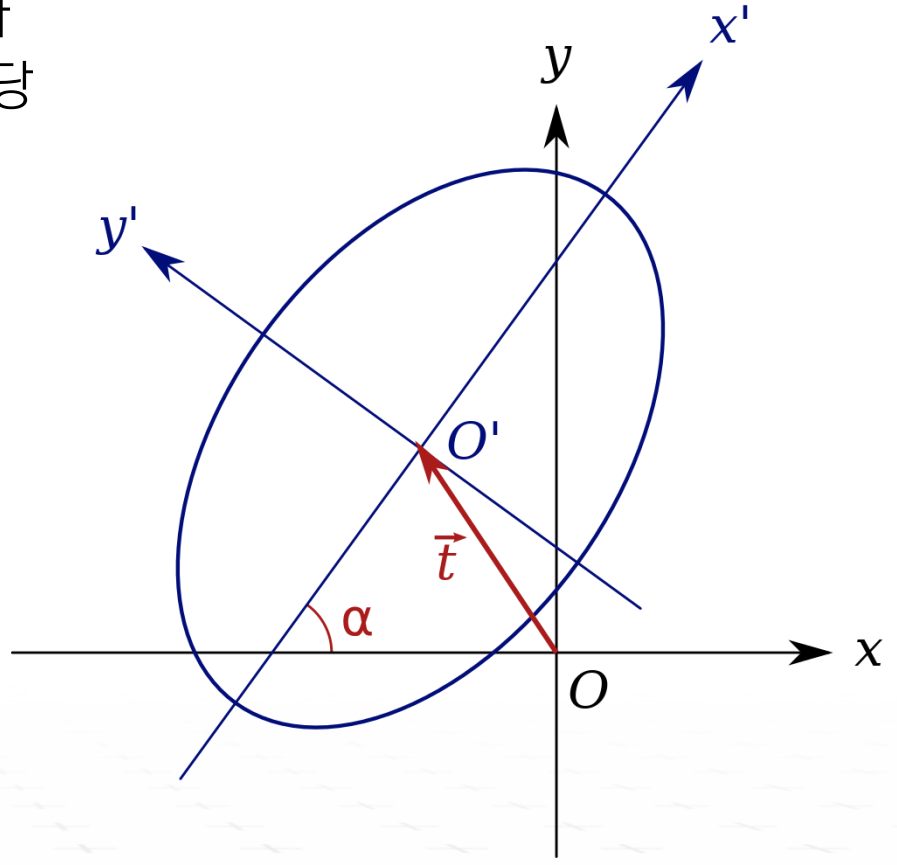
회전

$$R = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

반사

$$S = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$$

비틀기



1.활용

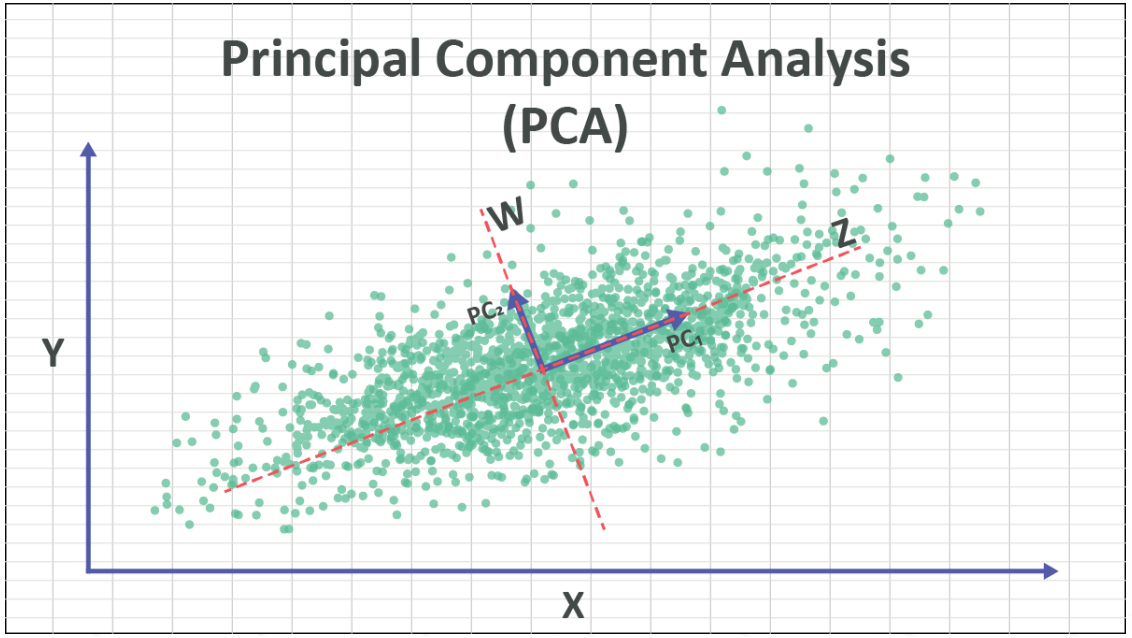
2) 주성분분석(Principal Component Analysis, PCA)

① 데이터의 분산을 최대한 유지하면서 중요한 특징을 찾는 기법

② 차원을 축소해주고 노이즈를 제거해주는 효과가 있음

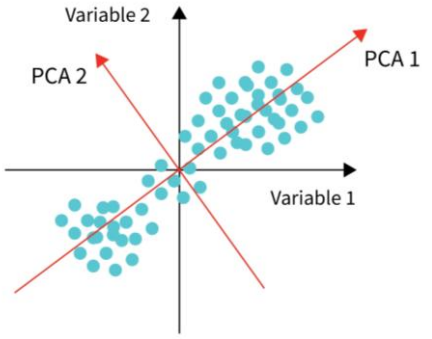
③ 주성분 분석 과정

- 공분산 행렬(상관계수)을 계산
- 공분산 행렬에 대한 고유값 분해를 수행
  - 고유값: **각 주성분이 설명하는 분산의 크기**
  - 고유벡터: 새로운 주성분의 방향
- 고유값을 기준으로 열을 재정렬한 후 주 성분을 선택



1.활용

2) 주성분분석(Principal Component Analysis, PCA)



500 x 3 행렬이 있다 가정

공분산이나 상관계수를 구하면 3x3으로 나옴

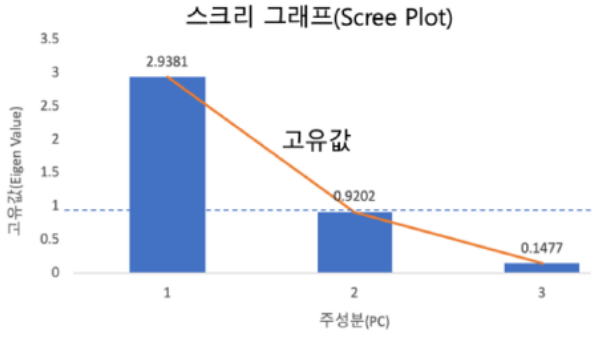
[1.00 0.85 -0.60
0.85 1.00 -0.40
-0.60 -0.40 1.00]

A변수와 B변수는 85% 양의 선형 상관관계가 있다.
A변수와 C변수는 60% 음의 선형 상관관계가 있다.
B변수와 C변수는 40% 음의 선형 상관관계가 있다.

행렬을 분해
- 이 때 다양한 방식을 쓸 수 있음.
고유값 분해, 특이값 분해(SVD), UV 분해 등등

$$\lambda_1 = 2.2523, \quad \lambda_2 = 0.6313, \quad \lambda_3 = 0.1164$$

$$\mathbf{v}_1 = \begin{bmatrix} 0.638 \\ 0.591 \\ -0.494 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0.170 \\ 0.518 \\ 0.838 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0.751 \\ -0.619 \\ 0.230 \end{bmatrix}$$



주축을 1개로 할꺼면
(500 x 3) 행렬을 v1(3 x 1)과 행렬곱 수행
(500 x 1)

주축을 2개로 할꺼면
(500 x 3) 행렬을 v1,v2 (3, 2)
(500 x 2)

Thank You

Email: qkdrk777777@naver.com