ADSP 3과목 Part2

통계 분석 개요

추출(sampling)
모집단(population) → 표본(sample)
모수(parameter) ← 통계량(statistic)
추론(inference)

모집단	- 잘 정의된 연구목적과 이와 연계된 명확한 연구대상(데이터 전체 집합) ex) 대통령 후보의 지지율-유권자
표본	- 모집단의 개체 수가 많아 전부 조사하기 힘들 때 모집단에서 추출 한 것 - 추출한 표본으로 모집단의 특성을 추론함(오차 발생) ex) 각종 여론조사에 참여한 유권자
통계량	- 표본의 특성을 나타내는 수치들
모수	- 모집단의 특성을 나타내는 수치들 - 모집단의 평균( ),분산( )같은 수치들을 모수라고함

확률적 표본추출법의 종류

단순 무작위 추출	- 모집단의 각 개체가 표본으로 선택될 확률이 동일하게추출되는 경우 - 모집단의 개체 수 N, 표본 수 n일 때 개별 개체가 선택될 확률은 n/N 임
계통추출	- 첫 번째 표본을 임의로 선택하고 일정 간격으로 다음 표본을 선택함 - 1~100번호 부여 후 , 1면 [1,11,21,31..91]선택
층화추출	- 모집단을 서로 겹치지 않게 몇 개의 집단 또는 층으로 나누고, 각 집단 내에서 원하는 크기의 표본을 단순 무작위추출법으로 추출함 - 층:성별,나이대,지역 등 차이가 존재하는 그룹
군집추출	- 모집단을 차이가 없는 여러 개의 집단(cluster)로 나눔 ex) 경상대학 내에 경영학과 경제학과

※ 비확률 표본 추출법은 특정 표본이 선정될 확률을 알 수 없어 통계학에서 사용할 수 없음

표본 추출 관련 오차의 종류 및 특징

표본오차/ 표본추출 오차	- 모집단을 대표하지 못하는 표본을 추출하여 발생하는 오차 - 모집단을 전수 조사하는 것이 아니라 표본을 추출하기 때문에 발생하는 오차 - 표본 오차는 표본의 크기가 커지만 작아짐, 전수조사에서는 0이됨
비표본추출 오차	- 표본 오차를 제외한 조사,집계,분석 과정에서 발생할 수 있는 모든 오차 ex)설문/측정 방식이 잘못됨,무응답/오류등 - 비표본 추출 오차는 표본의 크기에 비례하여 커짐, 표본의 크기가 크다고 반드시 좋은 것은 아님
표본 편의	- 표본 추출 과정에서 발생하는 편의(bias), 편의=추정값의 기댓값과 모수의 차이 ex)일반인으로부터 피험자를 모집했을 때,참여자들은 내용에 관심이 높은 사람일 가능성이 있어 표본이 치우칠 수 있음 - 확률화에 의해 최소화하거나 없앨 수 있음

척도의 종류

명목척도	- 단순히 측정 대상의 특성을 분류하거나 확인하기 위한 목적 - 숫자로 바꾸어도 그 값이 크고 작음을 나타내지 않고 범주를 표시함 ex) 성별,혈액형,출생지 등
서열(순위) 척도	- 대소 또는 높고 낮음 등의 순위만 제공할 뿐 양적인 비교는 X ex) 금,은,동메달,선호도 등
등간척도 (구간척도)	- 순위를 부여하되 순위 사이의 간격이 동일하여 양적인 비교 O - 절대 0점이 존재 X ex) 온도계 수치, 물가지수
비율척도	- 절대 0점이 존재하여 측정값 사이의 비율 계산이 가능한 척도 ex) 몸무게,나이,형제의 수 등

※ 절대 0점: 없음을 의미함(무)
- 온도의 0은 상대 0점으로 없음이 아니라 영상,영하의 중간 지점을 나타냄

※ 연속형 자료를 나타내는 척도로는 등간척도와 비율척도가 있다

 # 집중화 경향측정에 사용되는 값들

평균(Mean) : 이상치에 영향을 받음
중앙값(Median)
최빈값(Mode) : 이상치에 영향을 받지 않음

 # 데이터 집합이 얼마나 퍼져 있는지 알아보는데 사용하는 값들

산포도	- 자료의 변량들이 흩어져 있는 정도를 하나의 수로 나타낸 값 - 산포도가 크면 변량들이 평균으로부터 멀리흩어져 있음, 변동성이 커짐 - 산포도가 작으면 변량들이 평균 주위에 밀집, 변동성이 작아짐
편차	- 어떤 자료의 변량에서 평균을 뺀 값을 편차라고 한다. (편차 = 변량-평균)
분산(s2)	- 데이터 집합이 얼마나 퍼져 있는지 알아볼 수 있는 수치 - 편차의 제곱의 합을 n-1로 나눈것
표준편차(s)	- 자료의 산포도를 나타내는 수치,분산의 양의 제곱근 - 평균으로부터 각 데이터의 관찰 값까지의 평균거리

 # 분산,표준편차의 이해

- 툭정도시의 10가구를 표본으로 추출해 자녀수를 조사한 결과가 0,0,0,1,1,2,2,3,3,3 일 때
- 표본 평균 : 1.5 , 표본 분산 : 1.61 ,
  표본 표준편차 : 1.27 이 나옴
- 특정도시의 각 가구는 평균 1.5명의 자녀를 가지고, 각 가구는 약 1.27명의 자녀를 더하거나 뺀 범위 안에 있을 것으로 예상

※ 부산과 표준편차가 작을수록 자료들은 평균에 가까이 있음
※ 표준편차+평균+분산을 이용하면 변량을 구할수있음

 # 변동계수(CV)

- A학생이 평균 3시간 공부하고 표준편차는 0.4였고,B학생은 평균 6시간 공부하고 표준편차가 0.9이었다면 ㅇ떤 학생이 꾸준하게 공부했을까?
  CV=
- 이때, B학생의 표준편차가 0.8이라면 A,B학생의 변동계수가 같아짐.
- 관측되는 자료가 모두 양수일 때 사용

 # 범위(Range)

- **최소값과 최대값의 차이**
- **데이터가[1,3,5,7,10]인 경우 범위 → 9**

 # 통계 기본 용어

표본점	- ㅇ떤 행위를 했을 때 나올수 있는값 ex) 주사위 1.2.3.4.5.6 중 하나
표본공간	- 모든 표번점의 집합 ex) 주사위 굴리는 행위에 대한 표본공간S = [1,2,3,4,5,6]
사건	- 표본점의 특정한 집합 ex) 주사위를 한 번 굴렸을 때 홀수가 나오는 사건을 A라고하면 A =[1.3.5]
확률	- 사건이 일어날 수 있는 가능성을 수로 나타낸 것 - 어떤 사건을 A라고 했을 때, A가 발생할 확률은 P(A)로 표시 - 확률 = 사건/표본공간 - 확률값 = 0<=P(A)<=1

 # 사건의 종류

독립사건	- A의 발생이 B가 발생할 확률을 바꾸지 않는 사건 - 두 사건 A,B가 독립이면 ex) 주사위 던져서 나오는 눈의 값과 동전을 던져 나오는 앞/뒤 사건 ex) 서로 다른 사람이 총을 쏘아 과녁에 명중할 사건
배반사건	- 교집합이 공집합인 사건, 한쪽이 일어나면 다른 쪽이 일어나지 않을 때의 두 사건 ex) 동전 하나를 던져 앞면 나오는 사건, 뒷면 나오는 사건
종속사건	- 두 사건 A와 B에서 한 사건의 결과가 다른 사건에 영향을 주는 사건 ex) 음주와 사고 사건

 # 조건부확률

- 사건 B가 발생했다는 조건 아래서 사건 A가 발생할 조건부 확률

- 두 사건 A,B가 독립사건인 경우

ex) P(음주I사고)는 얼마인가?

	사고	무사고
음주자	0.07	0.23
비음주자	0.06	0.64

= (음주사고)/(음주사고+비음주사고)

= 0.07/0.13 = 0.54

 # 확률분포

분포	- 일정한 범위 안에 흩어져 퍼져 있는 정도
확률변수	- 확률현상: 어떤 결과들이 나올지 알지만, 가능한 결과들 중 어떤 결과가 나올지 모르는 현상
확률분포	- 어떤 확류변수가 취할 수 있는 모든 값들과 그 값을 취할 확률의 대응관계로 표시하는 것

확률변수

→ 동전을 2번 던질 때 앞면이 나온 횟수

확률분포는 다음과 같음(이산형 확률분포)

앞면횟수	0	1	2	합
확률	1/4	1/2	1/4	1

이산형 확률분포	- Discrete(별개의), 확률변수가 몇 개의 한정된 가능한 값을 가지는 분포 - 각 사건은 서로 독립이어야 함 ex) 이항분포/베르누이분포/기하분포/포아송분포 등
연속형 확률분포	- Continuous,확률변수의 가능한 값이 무한개이며 사실상 셀 수 없을 때 ex) 정규분포/지수분포/연속균일분포/카이제곱분포/F분포

 # 베르누이분포

- 싫험 결과 두 가지 중의 하나로 나오는 시행의 결과를 0 또는 1 값으로 대응시키는 확률변수 X에 대해 아래 식을 만족하는 확률변수 X가 따르는 확률분포

-

- 모수가 하나이며 서로 반복되는 사건이 일어나는 실험의 반복적 실행을 확률분포로 나타낸 것

베르누이분포의 예)

동전을 던져서 앞면이 나올 확률	p = 1/2 , q = 1/2
주사위를 던져서 4가 나올 확률	p = 1/6 , q = 5/6
주사위를 던져서 4,5가 나올확률	p = 1/3 , q = 2/3

 # 이항분포

- 서로 독립된 베르누이 시행을 n회 반복할 때 성공한 횟수를 x라 하면, 성공한 x의 확률분포를 말함
- 확률변수 K가 n,p 두 개의 모수를 갖으며, K~B(n,p)로 표시함
- n =1일 때 이항분포가 베르누이분포임
- 이항분포의 기댓값 : E(x) = np
- 이항분포의 분산 : V(x) = np(1-p)

이항분포의 예)

동전을 50번 던져서 앞면이 나올 경우는?	n = 50 , p = 1/2
주사위를 10번 던져서 나오는 눈이 5일 경우는?	n = 10 , p = 1/6
타율 3할인 타자가 100번 타석에 들어서면 안타를 얼마나 칠 것인가?	n = 100 , p = 0.3

 # 기하분포

- 베르누이 시행에서 처음 성공까지 시도한 횟수 x의 분포, 지지집합(x) = [1,2,3 ...]

 # 포아송분포

- 단위 시간이나 단위 공간에서 어떤 사건이 몇 번 발생할 것인지 표현하는 분포
- 특정 기간 동안 사건발생의 확률을 구할 때 사용

포아송분포의 예)

1. 어느 AS센터에 1시간당 평균 120건의 전화가 온다, 이때 1분 동안 걸려오는 전화 요청이 4건 이하일 확률은?
2. 어느 가게에 1시간당 평균 8명의 손님이 온다. 이때, 1시간 동안 손님이 10명 올 확률은?
3. 확률은 에서 최대이며, x가 커질수로 0에 근접

 # 기댓값 : 확률변수 X의 가능한 모든 값들의 가중평균

- 이산적 확률변수 기댓값
- 연속적 확률변수 기댓값

기댓값 예)
주사위 1개를 반복해서 던질 때 나타나는 기댓값
1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 3.5

 # 정규분포

-가우스 분포라고도 하며, 수집된 자료의 분포를 근사하는데 자주 사용함
-
- 평균0, 표준편차/분산 1인 정규 분포, N(0,1)를 표준정규 분포, z 분포라고 함
  예) 키,몸무게,시험 점수 등 거의 대부분의 측정값이 정규분포를 따름
- 정규분포의 평균 주위로 표준편차의 1배 범위에 있을 확률 68%, 2배 범위 안 95%, 3배 범위 안 99.7%

※ 확률 밀도 함수
- 특정 구간에 속할 확률을 계산하기 위한 함수

 # 대부분의 측정값을 정규분포로 가정하는 이유
"정규분포의 당위성"

이항분포의 근사	- 시행횟수 N이 커질 때,이항분포B(N,p)는 평균Np, 분산Npq인 정규분포와 N(Np,Npq)와 거의 같아짐
중심 극한 정리	- 표본의 크기가 N인 확률표본의 표본평균은 N이 충분히 크면 근사적으로 정규분포를 따르게 됨 - 모집단의 분포와 상관없이 표본의 크기가 30이상이 되면 N이 커짐에 따라 표본평균의 부포가 정규분포에 근사해짐
오차의 법칙	- - MLE : 실제 값일 가능성이 가장 높은 값 - 실제 값의 MLE가 측정값의 평균이라면, 오차는 정규분포를 따른다 → 오차의 법칙

 # 지수분포

- 사건이 서로 독립적일 때 다음 사건이 일어날 때가지 대기 시간은 지수분포를 따름
지수분포의 예)
전자 제품의 5년간 고장횟수가 평균 1회일 때, 1년 안에 고장 날 확률

 # t-분포

- 정규분포는 표본의 수가 적으면 신뢰도가 낮아짐 (n이 30개 미만인 경우)
- 표본을 많이 뽑지 못하는 경우에 대한 대응책으로 예측범위가 넓은 분포를 사용하며, 이것이 t-분포임
- t-분포는 표본의 개수에 따라 그래프의 모양이 변함 (표본의 개수가 많아질수록 정규분포와 비슷하며, 적을수록 옆으로 퍼짐)
- t-분포는 표본의 수가 30개 미만일 때 사용하며, '신뢰구간','가설검정'에 사용함

 # 카이제곱 분포

- 분산의 특징을 확률분포로 만든 것
- 카이제곱은 표준정규분포를 제곱한다는 의미가 내포
- 신뢰구간,가설검정에 사용하며, 그래프의 x축 좌표를 카이제곱값이라 부르며, 카이제곱분포표를 사용해 구하고 검정에 사용함
- 0이상의 값만 가질 수 있으며, 오른쪽 꼬리가 긴 비대칭모양
- 0의 오른쪽 부분에 분포가 많고, 0에서 멀어질수록 분포 감소
- 표본의 수가 많아지면 옆으로 넓적한 정규분포 형태가 됨

 # F분포

- 카이제곱분포와 같이 분산을 다룰 때 사용하는 분포
- 카이제곱분포는 한 집단의 분산, F분포는 두 집단의 분산을 다룸
- 두 집단의 분산이 크기가 서로 같은지 또는 다른지 비교하는데 사용함
- 카이제곱과 비슷하게 비대칭 모양이며, 양수만 존재

※ 언제 사용되는 분포일까?
ex) 한 집단 또는 두 두집단의 평균이 같은지를 검정
= z분포, t분포
ex) 한 집단의 모분산 검정(모수)
=
ex) 두 집단의 분산이 같은지를 검정
= F분포

 # 모집단에 대한 가정 여부에 따른 통계적 추론의 분류

모수적 추론 (Parametric Inference)	모집단에 특정 분포를 가정하고 모수에 대해 추론함
비모수적 추론 (Non-parametric Inference)	모집단에 대해 특정 분포 가정을 하지 않음

 # 추론 목적에 따른 통계적 추론의 분류

점추정	-하나의 값으로 모수의 값이 얼마인지 추측함 - 가장 참값이라고 여겨지는 하나의 모수의 값
구간 추정	- 모수를 포함할 것으로 기대되는 구간을 확률적으로 구함 - 일정한 크기의 신뢰수준으로 모수가 특정한 구간에 있을 것이라 선언하는 것

 # 표준오차(SE)

- 모집단에서 샘플을 무한 번 뽑아서 각 샘플마다 평균을 구했을 때, 그 평균들의 표준편차를 표준오차라 할 수 있음
- 표본평균이 모평균과 얼마나 떨어져 있는가를 나타냄
  n이 클수록 작은 값

 # 추정량

좋은 추정량 판단기준

일치성	- 표본의 크기가 커짐에 따라 표본 오차가 작아져야 한다.
비편향성,불편성	- 편향(bias) = 추정량의 기댓값 - 실제값(=모수의 값) - 추정량의 기댓값이 모수의 값과 같아야 한다 (편향 == 0)
효율성	- 추정량의 분산이 될 수 있는 대로 작아야 한다. (최소분산 추정량) - MSE가 작아야 한다.

 # 점추정

- 통계량 하나를 구하고 그것을 가지고 모수를 추정하는 방법
- '모수가 특정한 값일 것'이라고 추정하는 것
  (사실상 추정이 얼마나 정확한가 판단하기 불가능)

ex) A과목 수강 전체 학생 중 50명을 뽑아 조사한 결과 기말 점수가 80점 이었다면, 50명 뿐 아니라 나머지 A과목을 수강한 학생들의 점수도 80점 정도로 추정하는 것

※ 점추정량 구하는 법

1) 적률법 - 표본의 기댓값을 통해 모수를 추정
2) 최대가능도추정법(최대우도법) - 함수를 미분해서 기울기가 0인 위치에 존재하는 MLE를 찾는 방법
3) 최소제곱법 - 함수값과 측정값의 차이인 오차를 제곱한 합이 최소가 되는 함수를 구하는 방법

 # 구간추정

구간추정	- 점추정의 정확성을 보완하는 방법 - 통계량을 제시하는 것은 같지만 신뢰구간을 만들어서 추정하는 것
신뢰구간	- 모수가 포함되리라고 기대되는 '범위'
신뢰수준	- 모수값이 정해져 있을 때 다수 신뢰구간 중 모수값을 포함하는 신뢰구간이 존재할 확률 - 신뢰수준 95%의 의미 : n번 반복 추출하여 산정하는 신뢰구간들 중에서 평균적으로 95%는 모수 값을 포함하고 있을 것이라는 의미

ex1) 신뢰수준 95%에서 투표자의 35%~45%가 A후보를 지지하고 있다.

= 95%는 신뢰수준, 35%~45%는 신뢰구간이다.

ex2) 정치인 지지율 조사에서 A후보는 40%, B후보는 25%의 지지율을 얻었다. 신뢰수준 95%에서 표본오차는 3.1%포인트이다.

= 동일한 형태의 여론조사를 100번 실시했을 경우에 95번은 A후보가 40%에서 3.1%인 36.9%~43.1%, B후보는 25%에서 3.1%인 21.9%~28.1% 사이의 지지율을 얻을 것으로 기대된다는 의미이다.

※ 신뢰구간

 # 가설검정1

: 모집단에 대해 가설 설정 후, 표본관찰을 통해 그 가설의 채택 여부를 결정하는 통계적 추론 방법

추출(sampling)
모집단(population) → 표본(sample)
모수(parameter) ← 통계량(statistic)
추론(inference)

귀무가설	- 가설검정의 대상이 되는 가설 - 연구자가 부정하고자 하는 가설 - 효과 없음에 대한 가설
대립가설	- 귀무가설이 기각되면 채택되는 가설 - 연구자가 연구를 통해 입증/증명되기를 기대하는 예상이나 주장 - 효과 있음에 대한 가설

ex) 성적 관련 선생님의 가설
1) 귀무가설 (남학생과 여학생의 평균은 같다)
2) 대립가설 (남학생과 여학생의 평균은 다르다)

 # 가설검정2

제1종오류	- 귀무가설이 참인데 기각되는 오류 - 생산자 입장에서 정상제품을 불량품으로 판정하는 생산자 위험오류
제2종오류	- 귀무가설이 거짓인데 채택하는 오류 - 소비자 입장에서 불량품을 정상품으로 판정하는 소비자 위험오류

- 신뢰수준 : 1종오류를 범하지 않을 확률
- 검정력 : 2종오류를 범하지 않을 확률

※ 1) 두 가지 오류가 작을수록 바람직함
2)
3) 제1종 오류를 범할 확률의 최대 허용치를 특정값(유의수준)으로 지정해 놓고 제2종 오류의 확률을 가장 작게 해주는 검정 방법을 사용함

 # 가설검정3

기각역	- 귀무가설을 기각하고 대립가설을 채택하게 되는 영역 - 귀무가설이 옳다는 전제하에 구한 검정통계량의 분포에서 확률이 유의수준인 부분을 말한다.
유의수준	- 귀무가설이 참인데도 기각시키는 확률(제1종 오류 발생 확률)의 최대 허용한계 - 가능성이 '크다' 또는 '작다'의 판단기준 - 유의수준 0.05(5%) : 100번 실험에서 제1종 오류를 범하는 최대 허용 한계가 5번 - 유의수준 = 1-신뢰수준,유의수준 =

 # 가설점정4

유의확률(P-value)
- 귀무가설이 사실일 때 기각하는 1종 오류시, 우리가 내린 판정이 잘못되었을 확률
- 귀무가설의 신뢰구간을 벗어나는 확률
- 판정이 잘못되었을 확률
- P-value가 작을수록 그 정도가 약하다고 보며, P-value < 일 때, 귀무가설을 기각, 대립가설을 채택
- P-value가 0.05(5%) : 귀무가설을 기각했을 때 기각 결정이 잘못될 확률이 5%임

 # 모수적,비모수적 추론

모수적 추론	- 모집단에 특정 분포를 가정하고 분포의 특성을 결정하는 모수에 대해 추론하는 방법 - 모수로는 평균,분산등을 사용 - 자료가 정규분포,등간척도,비율척도인 경우 - ex) 온도의평균,몸무게의 표준

	편차 등
비모수적 추론	- 모집단에 대해 특정 분포 가정을 하지 않음 - 모수 자체보다 분포 형태에 고나한 검정을 실시함 - 표본 수가 적고, 명목척도,서열척도인 경우 ex) 성별,혈액형,두 그룹의 성비

※ 모수적 검정
1) 가정된 분포의 모수에 대해 가설 설정
2) 관측된 자료를 이용해 구한 표본 평균,표본 분산등을 이용해 검정 실시

※ 모수적 통계의 조건
- 표본의 모집단이 정규분포를 이루어야 하며, 집단 내의 분산은 같아야 함
- 변인(=변수)은 등간척도나 비율척도로 측정되어야 함 (아니면 비모수 통계 사용)

※ 모수 검정방법
: T test, ANOVA test, z분포, t분포 F분포,카이스퀘어 분포

 # T test
- 평균값이 올바른지, 두 집단의 평균 차이가 있는지를 검정하는 방법으로 t값을 사용함
- t값이 커질수록 p-value는 작아지며, 집단간 유의한 차이를 보일 가능성이 높아짐

t-검정 방법	예시
One Sample t-test	- 단일 표본의 평균 검정을 위한 방법 ex) S사 usb의 평균 수명은 20000 시간이다
Paired t-test	- 동일 개체에 어떤 처리를 하기 전, 후의 자료를 얻을 때 차이 값에 대한 평균 검정을 위한 방법 ex) 매일 1시간 한달 걸으면 2kg이 빠진다(걷기 전/수행 후) - 가능한 동일한 특성을 갖는 두 개체에 서로 다른 처리를 하여 그 처리의 효과를 비교하는 방법 ex) X질병 환자들을 두 집단으로 나누어 A,B 약을
Two sample t-test	- 서로 다른 두 그룹의 평균을 비교하여 두 표본의 차이가 있는지 검정하는 방법 - 귀무가설 - 두 집단의 평균 차이 값이 0이다 ex) 2학년과 3학년의결석률은 같다

One Sample t-test 예)
<임금의 평균이 100이다>

```
Wage=read.csv('../date/Wage.csv",
fileEncoding='UTF-8-BOM')
t.test(Wage$wage, mu=100)
```

One sample t-test

date : Wage$wage
t = 15.362, df = 2999,p-value < 2.2e-16
alternative hypothesis: true mean in not equlq to 100
95 percent confidence inferval :
110,2098 113.1974
sample estimates :
mean of x
111.7036

<u>해석</u>
1) df-2999 : n=df+1, 데이터의 수 3000개
2) 유의수준 5%에서 평균 wage=100 이라는 귀무 가설은 기각됨
3) 95% 신뢰구간 : 110.2098 ~ 113.1974
4) 귀무 가설에서 설정한 평균이 신뢰구간내에 존재하지 않음 (범위안에 평균=100 이없음!)

paired t-test (대응표본-t검정)의 예)
<수면유도제 데이터를 통한 '두 집단의 평균이 같다'는 가설>

t.test(extra~group, date=sleep, paired=TRUE)

paired t-test

data: extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences - 1.58

해석
1) paried=TRUE: Paired t-test, 짝을 이루는 데이터인 경우 분석 전 등분산성 검정 필요 없음
2) df = 9 : 그룹별 데이터의 수 10개
→ 분석 전 정규성 검정 실시
3) p-value가 0.002833으로 두 집단의 평균이 가다는 귀무가설이 기각할 수 있다.
4) 신뢰구간에 0이포함되지 않음

Two Sample t-test (=독립표본 t-test)의 예)
<수면유도제 데이터를 통한 '집단 간 평균이 같다'는 가설>

t.test(extra~group,date=sleep,var.equal=true)

Two sample t-test

data: extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.363874 0.203874
sample estimates:
mean in group 1 mean on group 2
0.75 2.33

해석
1) var.equal=TRUE: 두 집단의 분산이 같다는 등분산성 만족 → 분석 전 등분산성 검정 실시
2) df = 18 : 그룹이 2개이므로 데이터 수 20개
→ 분석 전 정규성 검정 실시
3) p-value가 0.07919로 두 집단의 평균이 같다는 귀무가설을 기각할 수 없다.
4) 신뢰구간에 0이 포함되므로 두 집단간 평균에 차이가 없다고 해석할 수 있음

 # 자유도 (df = n-1)
-통계적 추정에서 표본자료 중 모집단에 대한 정보를 주는 독립적인 자료의 수

 # 데이터 정규성 검정 종류 (Durbin-Watson X)

Q_Q plot	- 시각적으로 확인 하는 방법
Histogram	- 구간별 돗수를 그래프로 표시하여 시각적으로 정규분포를 확인하는 방법
Shapiro-Wilk test	- 귀무가설은 정규분포를 따른다로 p-value 가 0.05보다 크면 정규성을 가정하게 됨
Kolmogorov-Smirnov test	- K-S tsst, 두 모집단의 분포가 같은지 검정하는 것 - P-value가 0.05보다 크면 정규성을 가정하게 됨

 # 비모수적 추론

비모수적 검정	- 모집단의 분포에 대해 제약(정규분포,집단의 등분산 등)을 가하지 않고 실시하는 검정방법 - 평균,분산과 같은 모수 자체보다 분포 형태에 관한 검정을 실시함 - 모수적 방법보다 훨씬 단순함, 민감성을 잃을 수 있음 - 데이터의 개수가 작거나 범주형(명목,서열척도)데이터에 사용
비모수적 검정의 종류	- 명목척도 기준 : 카이스퀘어 검정 - 서열척도 기준 : Sign Test

 # 모수/비모수적 추론 방법

비교대상집단수	관계	비모수-명목척도	비모수-서열척도	모수
1		$\chi 2$적 합성 검정	kolmogorov-Smirnov test	One sample T test
2	독립	Crosstab $\chi 2$ 독립성검정	Mann-Whitney	Two sample T test
	대응자료	Mcnemar test	Wilcoxon signed-rank test Sign test	Paired T test
k(다변량)	독립	$\chi 2$동 질성 검정	Kruskal-Wallis test	ANOVA test
	대응자료	Cochran test	Friedman test	

ex) 다음 중 비모수적 추론이 아닌 것은?

 # 카이스퀘어

카이스퀘어 검정	- 한 개 범주형 변수와 각 그룹별 비율과 특정 상수비가 같은지 검정하는 적합도 검정 - 각 집단이 서로 유사한 성향을 갖는지 분석하는 동질성 검정 - 두 개 범주형 변수가 서로 독립이닞 검정하는 독립성 검정
예	1)적합도 검정 2)동질성 검정 3)독립성 검정

 # 부호검정

부호검정 (Sign test)	- 표본들이 서로 관련되어 있는 경우, 짝지어진 두 개의 관찰치들의 크고 작음을 +와 -로 표시하여 그 개수를 가지고 두 그룹의 분포 차이가 있는가에 대한 가설을 검증하는 방법
예	1)귀무가설 2)대립가설

 # 회귀분석

<용어정리>

독립변수	- 다른 변수에 영향을 받지 않고 독립적으로 변화하는 수, 설명변수라고도 함 - 입력 값이나 원인을 나타내는 변수 y=f(x)에서 x에 해당
종속변수	- 독립변수의 영향을 받아 값이 변화하는 수, 분석의 대상이 되는 변수 - 결과물이나 효과를 나타내는 변수 y=f(x)에서 y에 해당
잔차(오차항)	- 계산에 의해 얻어진 이론 값과 실제 관측이나 측정에 의해 얻어진 값의 차이 - 오차(Error) : 모집단 - 잔차(Residual) : 표본집단

※ 회귀분석
(ex, 영하 -5도에서는 오뎅이 몇 개나 팔릴까?)
- 변수와 변수 사이의 관계를 알아보기 위한 통계적 분석방법
- 독리변수의 값에의해 종속변수의 값을 예측하기 위함
- 일반 선형회귀는 종속변수가 연속형 변수일 때 가능
- 이산형(범주형) - 명목,서열척도
- 연속형 - 구간,비율척도

 # 회귀모형

- 선형회귀모형
한 개의 독립변수 : 단순 선형회귀
둘 이상의 독립변수 : 다중 선형회귀

<u>단일회귀 모형의 예)</u>

```
2 set. seed (2)
3 x = runif (50, 0, 5)
4 y = 5 + 2 * x + rnorm (50, 0, 0.5)
5 df <- data.frame (x,y)
6 fit <- lm(y~x, data=df)
7 fit

Call:
lm(formula = y ~ x, data = dfrm)

coefficients:
(Intercept)
4.748 → 절편  x 2.072 → 회귀계수
```

회귀방정식 : y=2.072*x + 4.748
runif(개수,시작,끝) : 시작~끝 범위에서 개수 만큼의 균일분포를 따르는 난수 발생
lm(y~x, data = df) : df에서 y를 종속변수, x를 독립변수로 회귀모형 생성

다중 회귀 모형의 예)

```
14 rm(list=is())
15 set.seed(10)
16 u <- runif (50, 0, 6)
17 v <- runif (50, 6, 12)
18 w <- runif (50, 3, 25)
19 y = 3 + 0.5 * u + 1* v - 2*w + rnorm (50, 0, 0.5)
20 df <- data. frame (y, u, v, w)

a <- lm(y~u+v+w, df)

Call:
lm(formula = y ~ u + v + w, data = df)

Coefficients:
(Intercept)        u        v        w
    3.4374   0.4676   0.9556  -1.9923
```

회귀방정식 : y=3.4374 + 0.4676*u + 0.9556*v - 1.9923*w

 # 최소자승법

- Y = aX + b 일 때 잔차를 제곱한 것의 합이 최소가 되도록 하는 상수 a.b를 찾는 것
- 큰 폭의 잔차에 대해 보가 더 큰 가중치를 부여하여, 독립변수 값이 동일한 평균치를 갖는 경우 가능한 변동 폭이 적은 표본회귀선을 도출하기 위한 것

 # 회귀 모형의 가정

- 선형성 : 독립변ㅂ수의 변화에 따라 종속변수도 변화하는 선형(linear)모형이다.
- 독립성 : 잔차와 독립변수의 값이 관련되어 있지 않다.
- 정규성 : 잔차항이 정규분포를 이뤄야 한다.
- 등분산성 : 잔차항들의 분포는 동일한 분산을 갖는다.
- 비상관성 : 잔차들끼리 상관이 없어야 한다.

Normal Q-Q plot	- 정규성(정상성), 잔차가 정규분포를 잘 따르고 있는지를 확인하는 그래프 - 잔차들이 그래프 선상에 있어야 이상적임
Scale-Location	- 등분산성, y축이 표준화 잔차를 나타내며, 기울기 0인 직선이 이상적임
Cook's Distance	- 일반적으로 1값이 넘어가면 관측치를 영향점(이상치)로 판별

 # Residuals vs Fitted는 선형성, 등분산성에 대해 알아볼 수 있는 그래프

- 선형성 : y값의 기울기가 0인 직선이 이상적
- 등분산성 : 점의 위치가 그래프에 고르게 분포하는 것이 이상적

 # 회귀모형 해석(평가방법)

- 표본 회귀선의 유의성 검정
 : 두 변수 사이에 선형관계가 성립하는지 검정하는 것
*

- 회귀모형 해석
 1)모형이 통계적으로 유의미한가?
 : F통계량,유의확률(p-value)로 확인
 2)회귀계수들이 유의미한가?
 : 회귀계수의 t값, 유의확률(p-value)로 확인

F통계량	- 모델의 통계적 유의성을 검정하기 위한 검정 통계량 - F통계량 = 회귀제곱평균(MSR) / 잔차제곱평균(MSE) - F통계량이 클수록 회귀모형은 통계적으로 유의미하다, p-value < 0.05 일 때 유의미함
결정계수	- 회귀식의 적합도를 재는 척도 - 결정계수는 0~1 사이의 범위를 갖음 - 결정계수가 커질수록 회귀방정식의 설명력이 높아짐

SST : Y의 변동성
SSE : X,Y를 통해 설명하지 못하는 변동성
SSR : Y를 설명하는 X의 변동성

회귀모형 해석(평가방법) 예)

```
a <- lm(y~u+v+w, df)
summary(a)

Call:
lm(formula = y ~ u + v + w, data = df)

Residuals:
     Min       10       Median     30       max
-1.06096  -0.31857  0.06092  0.32280  1.03220

Coefficients:
            Estimate Std. Error    t value   Pr(p>:t:)
(Intercept) 3.43742    0.46949   7.322   3.01e-09 ***
u           0.46762    0.04419 10.581   6.55e-14 ***
v           0.95558    0.04546   21.019   < 2e-16 ***
w       -1.99230     0.01052 -189.459 < 2e-16 ***
- - -
Signif. Codes: 0 “***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘*’
                0.1 ‘ ’ 1
Residual standard error: 0.4695 on 46 degrees of
freedom
Multiple R-squared: 0.9988
Adjusted R-squared:0.9987
F-statisticL; 1.254e+04 on 3 and 46 DFm
p-value: < 2.2e-16
```

해석

t통계량 = Estimate(회귀계수)/Std.Error(표준오차)
t통계량이 크다는 것은 표준오차가 작다는 의미
t통계량이 클수록 회귀계수가 유의하다.

- 다중회귀모형의 자유도(df) = n - k - 1
(n은 sample의 수, k는 독립변수의 수)

 # 다중공선성

- 모형의 일부 설명변수(=예측변수)가 다른 설명변수와 상관되어 있을 때 발생하는 조건
- 중대한 다중공선성은 회귀계수의 분산을 증가시켜 불안정하고 해석하기 어렵게 만들기 때문에 문제가 됨
- R의 vif함수를 사용해 구할 수 있으며, VIF값이 10이 넘으면 다중공선성이 존재한다고 봄
- 높은 상관 관계가 있는 설명변수를 모형에서 제거하는 것으로 해결해야함
- 설명변수를 제거하면 대부분 R-square가 감소함

※ 설명변수의 선택 원칙
- y에 영향을 미칠 수 있는 모든 설명변수 x들은 y의 값을 예측하는데 참여시킴
- 설명변수 x들의 수가 많아지면 관리에 많은 노력이 요구되므로 가능한 범위 내에서 적은 수의 설명변수를 포함시켜야 함
- 두 원칙이 이율배반적이므로 적절한 설명변수 선택이 필요함

 # 설명변수 선택 방법

모든 가능한 조합	- 모든 가능한 독립변수들의 조합에 대한 회귀모형을 고려해 AIC,BIC의 기준으로 가장 적합한 회귀 모형 선택 - AIC,BIC는 작은 값이 좋음
후진제거법	- 독립변수 후보 모두를 포함한 모형에서 출발해 제곱합의 기준으로 가장적은 영향을 주는 변수로부터 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하고, 이때 모형을 선택
전진선택법	절편만 있는 모델에서 출발해 기준 통계치를 가장 많이 개선시키는 변수를 차례로 추가하는 방법
단계별 선택법	모든 변수가 포함된 모델에서 출발해 기준 통계치에 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 빠져있는 변수 중에서 기준 통계치를 가장 개선시키는 변수를 추가함

 # 과적합

- 주어진 샘플들의 설명변수와 종속변수의 관계를 필요 이상 너무 자세하고 복잡하게 분석
- 샘플에 심취한 모델로 새로운 데이터가 주어졌을 때 제대로 예측해내기 어려울 수 있음
- 해결 방법으로 Feature(독립변수)의 개수를 줄이거나, Regularization(정규화)을 수행하는 방법이 있음

 # 정규화(Regularization)개념

- 베타값에 제약을 주어 모델에 변화를 주는 것
- λ값은 정규화 모형을 조정하는 hyper parameter
- λ값이 클수록 제약이 많아져 적은 변수가 사용되고, 해석이 쉬어지지만 underfitting 됨
- λ값이 작아질수록 제약이 적어 많은 변수가 사용되고.해석이 어려워지며 overfitting 됨

 # norm : 선형대수학에서 백터의 크기 또는 길이를 측정하는 방법

- L1 norm(=Manhattan norm) : 백터의 모든 성분의 절대값을 더함
- L2 norm(=Euclidean norm) : 출발점에서 도착점까지의 거리를 직선거리로 측정함

 # 라쏘(Lasso) 회귀 특징

- 변수 선택이 가능하며, 변수간 상관관계가 높으면 성능이 떨어짐
- L1 norm을 패널티를 가진 선형회귀방법, 회귀계수의 절대값이 클수록 패널티 부여
- w의 모든 원소가 0이 되거나 0에 가깝게 되게 해야 함 => 불필요 특성 제거

<장점>

- 제약 조건을 통해 일반화된 모형을 찾는다
- 가중치들이 0이 되게 함으로써 그에 해당하는 특성들을 제외해준다.

#Ridge 회귀 특성

- L2 norm을 사용해 패널티를 주는 방식
- 변수 선택이 불가능
- 변수간 상관관계가 높아도 좋은 성능을 가짐
- Lasso는 가중치들이 0이 되지만, Ridge의 가중치들은 0에 가까워질뿐 0이 되지는 않음

#엘라스틱넷 특성

- L1,L2 norm regularization
- 변수 선택 가능
- 변수 간 상관관계를 반영한 정규화가능

 # 상관계수의 이해

- 상관계수는 두 변수의 관련성의 정도를 의미함 (-1~1의 값으로 나타냄)
- 두 변수의 상관관계가 존재하지 않을 경우 상관계수는 '0'임
- 상관관계가 높다고 인과관계가 있다고 할수없음
- 피어슨 상관계수와 스피어만 상관계수가 있음
- 피어슨 상관계수 : 두 변수 간의 선형적인 크기만 측정가능
- 스피어만 상관계수 : 두 변수 간의 비선형적인 관계고 나타낼 수 있음
- R의 cor.test()함수를 사용해 상관계수 검정을 수행하고, 유의성검정을 판달할 수 있음
- 이때 귀무가설은'상관계수가 0이다'.
- 대립가설은'상관계수가 0이 아니다'

스피어만 상관계수	- 대상자료는 서열척도 사용, 두 변수 간의 비선형적인 관계를 나타낼 수 있음 - 연속형 외에 이산형도 가능. - **관계가 랜덤이거나 존재하지 않을 경우 상관 계수 모두 0에 가깝다** - 원시 데이터가 아니라 각 변수에 대해 순위를 매긴 값을 기반으로 함 - 두 변수 안의 순위가 완전 일치하면 1, 완전 반대면 -1 ex) 수학 잘하는 학생이 영어도 잘하는 것과 상관있는지 알아보는데 사용할 수 있음
피어슨 상관계수	- 대상자료는 등간척도, 비율척도 사용 - 두 변수 간의 선형적인 크기만 측정가능
공분산	- 2개의 확률변수의 선형 관계를 나타내는 값 - 하나의 변수가 상승하는 경향을 보일 때 다른 값도 상승하는 선형 상관성이 있다면 양의 공분산을 갖음 - 공분산이 0이면 서로 독립이며, 관측값들이 4면에 균일하게 분포되어 있다고 추정

상관분석의 예)
*귀무가설 : 상관계수가 0이다.

```
cor.test(c(1,3,5,7,9), c(1,2,4,6,8), method='pearson'
pearson's product-moment correlation

data: c(1,3,5,7,9) and c(1,2,4,6,8)
t = 15.588, df = 3, p-value = 0.0005737
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9065015 0.9996163
sample estimates:
      cor
0.9938837
```

 # 차원축소기법

1. 주성분분석
2. 요인분석
3. 판별분석
4. 군집분석
5. 정준상관분석
6. 다차원척도법

 # 다차원 척도법(MDS)

- 개체들 사이의 유사성,비유사성을 2차원 혹은 3차원 공간상에 점으로 표현하여 개체 사이의 군집을 시각적으로 표현하는 분석 방버
- 개체들의 거리는 유클리드(Euclidean) 거리와 유사도를 이용하여 구함
- 관측 대상의 상대적 거리의 정확도를 높이기 위해 적합 정도를 스트레스 값으로 나타내며, 0에 가⁇울수록 적합도가 좋음

 # 주성분분석(PCA)

- 데이터를 분석할 때 변수의 개수가 많다고 모두활용하는 것이 꼭 좋은 것은 아님
- 오히려 변수가 '다중공선성'이 있을 경우 분석 결과에 영향을 줄 수 있음
- 공분산행렬 또는 상관계수 행렬을 사용해 모든 변수들을 가장 잘 설명하는 주성분을 찾는 방법
- 상관관계가 있는 변수들을 선형 결합에 의해 상관관계가 없는 새로운 변수(주성분)을 만들고 분산을 극대화하는 변수로 축약함
- 주성분은 변수들의 선형결협으로 이루어져 있음
- 독립변수들과 주성분과의 거리인'정보손실량'을 최솨하하거나 분산을 최대화함

※ 공분산 행렬(default) vs 상관계수 행렬

- 공분산 행렬은 변수의 측정단위를 그대로 반영
- 상관계수 행렬은 모든 변수의 측정단위를 표준화함
- 공분산행렬을 이용한 경우 측정 단위를 그대로 반영하였기 때문에 변수들의 측정 단위에 민감
- 주성분 분석은 거리를 사용했기 때문에 척도에 영향받음 (정규화 전후의 결과가 다르다)
- 설문조사처럼 모든 변수들이 같은 수준으로 점수화된 경우 공분산행렬을 사용한다
- 변수들의 scale이 서로 많이 다른 경우에는 상관계수 행렬을 사용한다.

※ 주성분 분석에서 상관계수 행렬 사용
- prcomp(data, scale=TRUE)
- princomp(data, cor=TRUE)

 # 시계열 자료

시계열 자료	- 시간의 흐름에 따라 관측된 데이터 - 시계열 분석을 위해서는 정상성을 만족해야 함
정상성	- 시계열의 수준과 분산에 체계적인 변화가 없고, 주기적 변동이 없다는 것 - 미래는 확률적으로 과거와 동일하다는 것
정상 시계열의 조건	- 평균은 모든 시점(시간t)에 대해 일정하다 - 분산은 모든 시점(시간t)에 대해 일정하다 - 공분산은 시점(시간t)에 의존하지 않고, 단지 시차에만 의존한다

 # 정상시계열로 전환하는 방법

평균이 일정하지 않은 경우	원계열에 차분 사용
계절성을 갖는 비정상시계열	계절 차분 사용
분산이 일정하지 않은 경우	원계열에 자연로그(변환)사용

차분 : 현 시점의 자료 값에서 전 시점의 자료 값을 빼주는 것을 의미

 # 시계열 모형

AR모형 자기회귀모형	- AR(P) : 현 시점의 자료를 P시점 전의 유한개의 자기 자신의 과거 값을 사용하여 설명 - 백색 잡음의 현재 값과 자기 자신의 과거 값의 선형 가중 값으로 이루어진 정상 확률 모형
MA모형 이동평균 모형	- MA(q) : 과거 q시점 이전 오차들에서 현재 항의 상태를 추론한다 - 각 과거치는 동일 가중치가 주어짐 - 현시점의 자료가 유한개의 과거 백색잡음의 선형결합으로 표현되었기 때문에 항상 정상성을 만족함
ARIMA모형	- 현재와 추세간의 관계를 정의한 것 - 많은 시계열 자료가 ARIMA모형을 따름 - 비정상시계열 모형 - 차분/변환을 통해 AR,MA,ARMA 모형으로 정상화 할 수 있음 - ARIMA(p,d,q) p:AR모형차수 d:차분 q:MA모형 차수

ex) ARIMA(1,2,3)이라면 2번 차분해서 ARMA모형이 될 수 있음

ex) ARIMA(0,1,3) : IMA(1,3)모형이고 이것을 1번 차분하면 MA(3) 모형이됨

ex) ARIMA(2,3,0) : ARI(2,3)모형이고, 이것을 3번 차분하면 AR(2) 모형이됨

 # 분해시계열

: 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법

추세요인	자료의 그림을 그렸을 때 그 형태가 오르거나 내리는 등 자료가 어떤 특정한 형태를 취할 때
계절요인	계절에따라, 고정된 주기에 따라 자료가 변화하는 경우
순환요인	알려지지 않은 주기를 가지고 자료가 변화하는 경우
불규칙요인	위 세 가지 요인으로 설명할 수 없는 회귀분석에서 오차에 해당하는 요인에 의해 발생하는 경우

ADSP 3과목 Part3

 # 데이터마이닝

: 모든 사용가능한 원천 데이터를 기반으로 감춰진지식, 기대하지 못했던 경향 또는 새로운 규칙 등을 발견하고 이를 실제 비즈니스 의사결정 등에 유용한정보로 활용하는 일련의 작업

※ 데이터 마이닝 5단계

1. 목적 정의 : 데이터 마이닝 도입 목적을 명확
2. 데이터 준비 : 데이터 경제를 통해 데이터의 품질 확보까지 포함, 필요시 데이터 양 충분하게 확보
3. 데이터 가공 : 목적 변수를 정의하고, 필요한 데이터를 데이터 마이닝 소프트웨어에 적용할 수 있게 가공 및 준비하는 단계
4. 데이터 마이닝 기법 적용
5. 검증

 # 데이터 마이닝 기법

분류	- 기존의 분류, 정의된 집합에 배정하는 것
추정	- 알려지지 않은 결과의 값을 추정하는 것
연관분석	- '같이 팔리는 물건' 같이 아이템의 연관성을 파악하는 분석 - 카탈로그 배열 및 교차판매 - 공격적 판촉행사 등의 마케팅 계획에 사용
예측	- 미래에 대한 것을 예측
군집	- 미리 정의된 기준이나 예시에 의해서가 아닌 레코드 자체가 가진 다른 레코드와의 유사성에 의해 그룹화되고 이질성에 의해 세분화됨
기술	- 데이터가 가진 특징 및 의미를 단순하게 설명하는 것

 # 로지스틱 회귀분석

- 독립변수는 연속형, 종속변수가 범주형인 경우 적용되는 회귀분석 모형
- 종속변수가 성공/실패, 사망/생존과 같이 이항변수(0,1)로 되어 있을 때 종속변수와 독립변수 간의 관계식을 이용하여 두 집단 또는 그 이상의 집단을 분류하고자 할 때 사용하는 분석기법

	일반 성형 회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	이산형(범주형) 변수
모형 탐색 방법	최 소 자 승 법 (LSM,최소제곱법)	최 대 우 도 법 (MLE),가중최소자승법
모형 검정	F-test,T-test	x2 test

※ sigmoid : Logistic 함수라 불리기도 하며, 비선형적 값을 얻기 위해 사용

※ 회귀식에 대한 해석 방법이 선형회귀와 다름

```
2 prob <- 0.8
3 odds <- prob / (1-prob)
4 log_odds <- log(odds)
5 r <- 1 / (1 + exp(-log_odds))
```

values	
log_odds	1.38629436111989
odds	4
prob	0.8
r	0.8

해석)

승산(odds) = 성공률/실패율, Pi/(1-P) 단, Pi=성공률

- 성공이 일어날 가능성이 높은 경우는 1.0 보다 큰 값
- 실패가 발생할 가능성이 높은 경우는 1.0 보다작은값
- 확률에 대해 0~무한값으로 변환한 값

log odds, logit transformation = log(odds)

- 선형화의 하나로, odds값에 log를 취하여 값의 범위를 전체 실수 범위로 확장함

sigmoid 함수

- log_odds 값을 연속형 0~1 사이의 값으로 바꾸는 함수
- 비선형 값을 얻기 위해 사용

※ 로지스틱 회귀분석 해석

```
8 prob_a <- 0.5
9 prob_b <- 0.2
10 odds_a <- prob_a / (1-prob_a)
11 odds_b <- prob_b / (1-prob_b)
12 odds_ratio <- odds_a / odds_b
```

Values	
odds_a	1
odds_b	0.25
odds_ratio	4

승산비(odds ratio) = 관심있는 사건이 발생할 상대 비율, x=1일 때, y=1이 되는 상대적 비율

- 로지스틱 회귀에서 exp(x1)의 의미(단,x1:회귀계수)
- 나머지 변수가 주어질 때x1이 한 단위 증가할 때마다 성공(Y=1)의 odds가 몇 배 증가하는지를 나타냄

의사 결정 나무모형

- 의사 결정 규칙을 나무 구조로 나타내 전체 자료를 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석 방법
- 분석과정이 직관적이고 이해하기 쉬움

특징	- 새로운 데이터를 분류하거나 값을 예측하는 것 - 분리 변수 p차원 공간에 대한 현재 분할은 이전 분할에 영향을 받는다. - 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나간다(불순도 감소)
종류	- 목표변수(=종속변수)가 이산형인 경우의 분류나무 - 목표변수가 연속형인 경우의 회귀나무
장점	- 구조가 단순하여 해석이 용이함 - 비모수적 모형으로 선형성,정규성,등분산성 등의 수학적 가정이 필요없음 - 범주형(이산형)과 수치형(연속형)변수를 모두 사용할 수 있음
단점	- 분류 기준값의 경계선 부근의 자료값에 대해서는 오차가 큼(비연속성) - 로지스틱회귀와 같이 각 예측변수의 효과를 파악하기 어려움 - 새로운 자료에 대한 예측이 불안정할 수 있음

※ 독립변수 : 설명변수/예측변수/Feature
※ 종속변수 : 목표변수/반응변수/Label

의사결정나무의 결정규칙

분리기준	- 순수도가 높아지는 방향으로 분리 - 불확실성이 낮아지는 방향
정지규칙	- 더 이상 분리가 일어나지 않고 현재의 마디가 최종마디가 되도록 하는 규칙 - '불순도 감소량'이 아주 작을 때 정지함
가지치기 규칙	- 최종 노드가 너무 많으면 Overfitting 가능성이 커짐, 이를 해결하기위해 사용 - 가지치기의 비용함수을 최소로 하는 분기를 찾아내도록 학습 - 별도 규칙을 제공하거나 경험에 의해 실행할 수 있음

 # 불순도 측정 지표

: 목표변수가 범주형일 때 사용하는 지표

지니 지수	- 불순도 측정지표, 값이 작을수록 순수도가 높음(분류 잘됨 -
엔트로피 지수	- 불순도 측정지표, 가장 작은 값을 갖는 방법 선택
카이제곱 통계량의 유의 확률(p-value)	- 가장 작은 값을 갖는 방법 선택

 # 의사결정나무를 위한 알고리즘

: 의사결정나무를 위한 알고리즘은 CHAID,CART,ID2, C5.0,C4.5가 있으며 **하향식 접근 방법**을 이용

※ 알고리즘 별 분리,정지 기준변수 선택법

알고리즘	이산형 목표변수(분류나무)	연속형 목표변수(회귀나무)
CART (Classification And Regression Tree)	지니지수	분산 감소량
C5.0	엔트로피지수	
CHAID (Chi-squared Automatic interaction Detection)	카이제곱 통계량의 P-value	ANOVA F-통계량-P-value

<u>의사결정 트리 예)</u>

```
2 # CART 알고리즘
3 library(rpart)
4 a <- rpart(species~. , data=iris)
5 a
6 plot(a, compress=T, margin=.3)
7 text(a, cex=1)

9 install.packages('rpart.plot')
10 library(rpart.plot)
11 prp(a, type=4, extra=2, digits=3)
```

```
n= 150

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 150 100 setosa (0.33333 0.33333 0.33333)
2) petal.Length< 2.45 50 0 setosa (1.0000 0.0000
0.0000) *
3) petal.Length>= 2.45 100 50setosa (1.0000
0.0000 0.0000) *
```

 # 앙상블 모형

- 여러 개의 분류 모형에 의한 결과를 종합하여 분류의 정확도를 높이는 모형
- 약하게 학습 된 여러 모델들을 결합하여 사용
- 성능을 분산시키기 때문에 과적합 감소효과가 있음

<종류>

Voting
Bagging
Boosting
Random Forest

※ Voting

- 서로 다른 여러 개 알고리즘 분류기 사용
- 각 모델의 결과를 취합하여 많은 결과 또는 높은 확률로 나온 것을 최종 결과로 채택하는 것

※ **배깅(Bagging)**
- 서로 다른 훈련 데이터 샘플로 훈련, 서로 같은 알고리즘 분류기 결합
- 원 데이터에서 중복을 허용하는 크기가 같은 표본을 여러 번 단순 임의 복원추출하여 각 표본에 대해 분류기를 생성하는 기법
- 여러 모델이 병렬로 학습, 그 결과를 집계하는 방식
- 같은 데이터가 여러 번 추출될 수도 있고, 어떤 데이터는 추출되지 않을 수 있음
- 데이터 집합으로부터 크기가 같은 표본을 여러번 단순 임의 복원 추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 방법

※부스팅(Boosting)
- 여러 모델이 순차적으로 학습
- 재표본 과정에서 각 자료에 동일한 확률을 부여하지 않고, 분류가 잘못된 데이터에 더 가중을 주어 표본을 추출하는 분석방법
- 맞추기 어려운 문제를 맞추는데 초점이 맞춰져있고, 이상치(outlier)에 약함
- 대표적 알고리즘 : Light GMB (Leaf-wise-node 방법을 사용하는 알고리즘)

※ **랜덤 포레스트(Random forest)**
- 배깅에 랜덤 과정을 추가한 방법
- 설명변수의 일부분만을 고려함으로 성능을 높이는 방법 사용
- 여러 개 의사결정 나무를 사용해, 하나의 나무를 사용할 때보다 과적합 문제를 피할 수 있음

 # k-NN
- 새로운 데이터에 대해 주어진 이웃의 개수(k)만큼 가까운 멤버들과 비교하여 결과를 판단하는 방법
- k값에 따라 소속되는 그룹이 달라질 수 있음 (k값은 hyper parameter)
- 거리를 측정해 이웃들을 뽑기 때문에 스케일링이 중요함
- 모형을 미리 만들지 않고, 새로운 데이터가 들어오면 그때부터 계산을 시작하는 lazy learning이 사용되는 지도학습 알고리즘

 # SVM
- 아래 그림에서 H3는 분류를 올바르게 하지 못하며, H1, H2는 분류를 올바르게 하는데 H1가 H1보다 더 큰 간격을 갖고 분류하므로 이것이 분류 기준이 됨

 # 인공 신경망(ANN) 모형
- 인공신경망을 이용하면 분류 및 예측을 할 수 있음.
- 분석가의 주관과 경험에 따른다.
- 입력층,은닉층,출력층 3개의 층으로 구성되어 있고, 각 층에 뉴런이 여러 개 포함되어 있음
- 학습 : 입력에 대한 올바른 출력이 나오도록 가중치(weights)를 조절하는 것
- bias, variance : 학습 알고리즘이 갖는 두 가지 종류의 error로 trade off관계임

※ bias : 지나치게 단순한 모델로 인한 error, bias가 크면 과소 적합을 야기함

※ variance : 지나치게 복잡한 모델로 인한 error, cariance가 크면 과대 적합이 야기됨

※ 학습 모형이 유연하다는 것은 복잡도가 증가한다는 것을 의미 (bias 낮고, variance가 높음)

 # 경사하강법

- 함수 기울기를 낮은 쪽으로 계속 이동시켜 극값에이를 때까지 반복시키는 것
- 제시된 함수의 기울기의 최소값을 찾아내는 머신러닝 알고리즘
- 비용함수를 최소화 하기 위해 parameter를 반복적으로 조정하는 과정

 # 인공 신경망 모형의 장/단점

장점	- 복잡한 비선형 관계에 유용 - 이상치 잡음에 대해서도 민감하게 반응하지 않음 - 입력변수와 결과변수가 연속형이나 이산형인 경우 모두 처리 가능
단점	- 결과에 대한 해석이 쉽지 않음 - 최적의 모형을 도출하는 것이 상대적으로 어려움 - 데이터를 정규화 하지 않으면 지역해(local minimum)에 빠질 위험이 있음

 # 신경망 활성화 함수

- 결과값을 내보낼 때 사용하는 함수로, 가중치 값을 학습할 때 에러가 적게 나도록 도움
- 풀고자 하는 문제 종류에 따라 활성화 함수의 선택이 달라짐

※ 활성화 함수의 종류

 # 신경망 활성화 함수

sigmoid 함수	- 연속형 0~1, Logistic함수라 불리기도 함 - 선형적인 멀티-퍼셉트론에서 비선형 값을 얻기 위해 사용
softmax 함수	- 모든 logits의 합이 1이 되도록 output을 정규화 - 주로 3개이상 분류시 사용함 - sigmoid 함수의 일반화된 형태로 목표치가 다 범주인 경우 각 범주에 속할 사후 확률을 제공하는 활성화 함수

 # 신경망 은닉 층, 은닉 노드

- 다층신경망은 단층신경망에 비해 훈련이 어려움
- 은닉층 수와 은닉 노드 수의 결정은 '분석가가 분석 경험에 의해 설정'함

은닉 층 노드가 너무 적으면	- 네트워크가 복잡한 의사결정 경계를 만들 수 없음 - unerfitting 문제 발생
은닉 층 노드가 너무 많으면	- 복잡성을 잡아낼 수있지만, 일반화가 어렵다. - 레이어가 많아지면 기울기 소실 문제가 발생할 수 있다 - 과적합(Overfitting)문제 발생
역전파 알고리즘	- 출력층에서 제시한 값에 대해, 실제 원하는 값으로 학습하는 방법으로 사용 - 동일 입력층에 대해 원하는 값이 출력되도록 개개의 weight를 조정하는 방법으로 사용됨
기울기 소실 문제	- 다층신경망에서 은닉층이 많아 인공신경망 기울기 값을 베이스로하는 역전파 알고리즘으로 학습시키려고 할 때 발생하는 문제

 # 기울기 소실

- 다층신경망에서는 역전파 알고리즘이 입력층으로 갈수록 Gradient가 점차적으로 작아져 0에 수렴하여, weight가 업데이트 되지 않는 현상
- 은닉층이 늘어나면서 기울기가 중간에 0이 되어 버리는 문제

 # 모형 평가

홀드아웃	- 원천 데이터를 랜덤하게 두 분류로 분리하여 교차검정을 실시하는 방법으로 하나는 모형 학습 및 구축을 위한 훈련용 자료로 다른 하나는 성과평가를 위한 검증용 자료로 사용하는 방법 - 과적합 발생 여부를 확인하기 위해서 주어진 데이터의 일정 부분을 모델을 만드는 훈련데이터로 사용하고, 나머지 데이터를 사용해 모델을 평가 - 2종 오류의 발생을 방지 - iris데이터를 7:3 비율로 나누어 Training에서 70%, Testing에 30% 사용하도록 하는 것
교차검증	- 데이터가 충분하지 않을 경우 Hold-out으로 나누면 많은 양의 분산 발생 - 이에 대한 해결책으로 교차검증을 사용할 수 있음, 그러나 클래스 불균형 데이터에는 적합하지않음 - 주어진 데이터를 가지고 반복적으로 성과를 측정하여 그 결과를 평균한 것으로 분류 분석 모형의 평가 방법
붓스트랩	- 평가를 반복하는 측면에서 교차검증과 유사하지만, 훈련용 자료를 반복 재선정한다는 점에서 차이 - 관측치를 한 번 이상 훈련용 자료로 사용하는 복원추출법에 기반 - 훈련 데이터를 63.2% 사용하는 0.632 붓스트랩이 있음 - 반복 수행 시 매회 다른 데이터 분할이 된다.

 # 데이터 분할 시 고려사항

- class의 비율이 한쪽에 치우쳐 있는 클래스 불균형 상태라면 다음 기법 사용을 고려한다.

1) under sampling : 적은 class의 수에 맞추는 것
2) over sampling : 많은 class의 수에 맞추는 것

※ 훈련 데이터에 대한 학습만을 바탕으로 모델의 설정(Hyperparameter)를 튜닝하게 되면 과대적합이 일어날 가능성이 매우 크다.
※ tsst set결과가 일반적으로 training set 결과보다 좋지 않다.

 # 오분류표를 활용한 평가 지표

T/F	P/N
실제 == 예측 : TRUE	TRUE 예측 : Positive
실제 ! = 예측 : FALSE	FALSE 예측 : Negative

TP = positive로 예측해서 맞춘 것
FP = 예측을 Positive로 했는데 틀림(=negative)

구분		실제	
		FALSE	TRUE
예측	FALSE	TN	**FN(2종)**
	TRUE	**FP(1종)**	TP

※ **Pricision**(정밀도)
(**예측**값이 TRUE인것에 대해 실제 값이 TRUE인 지표)
= TP/(TP+FP)

※ **Error Rate**
(전체 예측에서 틀린 예측의 비율)
= (FP+FN) / (TP_TN+FP+FN)

※ **Sensitivity, Recall**
(**실제** 값이 TRUE인것에 대해 예측 값이 TRUE인 지표)
= TP/(TP+FN)

※ **Accuracy**
(전체 예측에서 옳은 예측의 비율)
(불균형한 레이블 값 분포의 데이터에서는 모델의 성능이 실제로 좋지 못하더라도 정확도가 높을 수 있음)
= (TP+TN) / (TP+TN+FP+FN)

※ **Specificity** (특이도)
= TN / (TN+FP)

※ **F1 Score**
(불균형한 데이터 평가에 사용)
= 2*(Precision*Recall) / (Precision+Recall)

 # 카파 상관계수(kappa)

- 코헨(Cohen)의 상관계수로 두 평가자의 평가가 얼마나 일치하는지 평가하는 값

 # ROC Curve

- ROC 그래프의 밑부분의 면적이 넓을수록 좋은 모형으로 평가함
- FP-Ratio(1-특이도), 민감도를 나타내어 이 두 평면값의 관계로 하는 모형평가
- **FP-Rate** : FP/(FP+TN)
- 1-Specificity : 실제가 FALSE인데 예측이 TRUE로 된 비율(1종오류비율)

※ Perfect classifier

: 긍정,부정 모두 다 맞추는 위치로 classification성능이 우수하다고 봄, (**X=0, Y=1인 경우**)

 # 이익 도표

- 분류 모형의 예측 성능을 평가하기 위한 척도, 주로 불균형 데이터 집합에 사용됨
- 랜덤 모델과 비교하여 해당 분류 모델의 성과가 얼마나 향상되었는지각 등급별로 파악할 수 있음
- 정보를 산출하여 나타내는 표

 # 향상도 차트

- 좋은모델 : Lift Curve가 빠른 속도로 감소 추세를 보임

 # 군집분석

- 여러 변수 값들로부터 n개의 개체를 유사한 성격을 가지는 몇 개의 군집으로 집단화하고 형성된 군집들의 특성을 파악해 군집들 사이의 관계를 분석하는 다변량분석 기법

계층적 군집	**응집형** : 단일(최단)연결법,ward 연결법
	분리형 : 다이아나 방법
분할적 군집	프로토타입-기반 - K-중심 군집 : k-평균 군집, k-중앙값 군집, k-메도이드 군집
	분포기반 - 혼합 분포 군집
	밀도기반 - 중심밀도 군집,격자기반 군집

 # 계층적 군집 분석의 특징

- 가장 유사한 개체를 묶어 나가는 과정을 반복하여 원하는 개수의 군집으 형성하는 방법
- 유사도 판단은 두 개체 간의 거리에 기반하므로 거리 측정에 대한 정의가 필요함
- 이상치에 민감함
- 사전에 군집 수 **k를 설정할 필요가 없는** 탐색적 모형
- 병합적 방법에서 한 번 군집이 형성되면 군집에 속한 개체는 다른 군집으로 이동할 수 없음

 # 계층적 군집 - 응집형(병합군집)군집방법

최단연결법	- **단일연결법**이라고도 하며, 두 군집 사이의 거리를 군집에서 하나씩 관측 값을 뽑았을 때 나타날 수 있는 거리의 **최솟값**을 추정
최장연결법	- **완전연결법**이라고도 하며, 거리의 **최대값**을 측정
중심 연결법	- 두 군집의 중심 간의 거리를 측정
와드 연결법	- 계층적 군집내의 오차제곱합에 기초하여 군집을 수행하는 군집방법
평균 연결법	- 계산양이 많아질 수 있음

 # 계층적 군집의 거리

수학적 거리 개념 : 유클리드,맨해튼,민코프스키

통계적 거리 개념 : 표준화, 마할라노비스

유클리드	- 두 점 사이의 가장 직관적이고 일반적인 거리의 개념, - 방향성이 고려되지 않음
맨해튼	- 두 점의 각 성분별 차의 절대값 합
민코프스키	- q=2이면 유클리드 - q=1이면 맨해튼
표준화 거리	- 각 변수를 해당 변수의 표준편차로 척도 변환한 후 유클리드 거리를 계산한 것으로 통계적 거리라고함
마할라노비스	- 변수의 표준화와 함께 변수 간의 상관성을 동시에 고려한 통계적 거리
dist 함수	- 거리측정에 사용하는 함수로 사용가능한 거리 개념으로 유클리드,맨해튼,민코프스키,Maximum,canberra,binary 등이 있음
코사인 거리	- 두 벡터 사이의 사잇각을 계산해서 유사한 정도를 구하는 것

 # 비계층적 군집 - 분할적 군집 방법 (k-중심 군집)

k-means : k-mean 방법은 사전에 군집의 수k를 정해 주어야 함 (k:hyper-parameter)

k-means 절차

1)초기 군집의 중심으로 k개의 객체를 임의로 선택한다

2)각 자료를 가장 가ᄀᆞ운 군집의 중심에 할당한다

3)각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다.

4)군집 중심의 변화가 거의 없을 때까지 2,3을 반복

 # 비계층적군집

DBSCAN	- 밀도 기반 클러스터링으로 점이 세밀하게 몰려 있어 밀도가 높은 부분을 클러스터링 함 - 어느 점을 기준으로 반경 내에 점이 n개 이상 있으면 하나의 군집으로 인식 - 임의적 모양의 군집분석에 적합 - k 값을 정할 필요 없음 - 이상치(outlier)에 의한 성능 하락을 완화할 수 있음
혼합분포군집	
EM 알고리즘	1. 모수(평균,분산,혼합계수)에 대해 임의의 초기값을 정함 2. E step : k개의 모형 군집에 대해 모수를 사용해 각 군집에 속할 사후 확률을 구함 3. M step : 사후확률을 이용해 최대 우도 추정으로 모수를 다시 추정하고, 이를 반복함

 # 실루엣 계수

- 군집분석에서 중요한 지표로서,거리가 가까울수록 높고 멀수록 낮은 지표이자 완벽히 분리된 경우 1이 되는 지표
- 군집내 거리와 군집 간의 거리를 기준으로 군집 분할 성과를 측정하는 방식
- 클러스터 안의 데이터들이 다른 클러스터와 비교해 얼마나 비슷한가를 나타내는 군집평가
- 실루엣 지표가 1에 가까울수록 군집화가 잘 되었다고 판단

※ 군집 결과에 대한 안정성 검토는 실루엣,Dunn Index를 사용함

※ Height 150에서 선을 그어 만나는 선의 수가 군집의 수이다.

 # SOM (Self-Organizing Maps)

- 자기조직화지도
- 인공신경망의 한 종류로, 차원축소와 군집화를 동시에 수행하는 기법
- 비지도 학습
- 고차원으로 표현된 데이터를 저차원으로 변환해서 보는데 유용함
- 주요 기능 중에 데이터의 특징을 파악하여 유사 데이터를 클러스터링한다.
- 2개의 층으로 구성 (다층x)

 # SOM Process

1단계 : SOM의 노드에 대한 연결강도(weight)초기화

2단계 : 입력 벡터와 경쟁 층 노드 간의 거리 계산 및 입력벡터와 가까운 노드 선택 → 경쟁

3단계 : 경쟁에서 선택된 노드와 이웃 노드의 가중치(연결강도)갱신 → 협력 및 적응

4단계 : 단계 2로 가서 반복

- 승자만이 출력을 내고, 승자와 그의 이웃만이 연결강도를 수정하는 승자 독점 구조로 인해 경쟁층에는 승자 뉴련만 나타남

 # 신경망 모형 vs SOM

신경망 모형	- 연속적인 layer로 구성 - 에러 수정을 학습 - 역전파 알고리즘
SOM	- 2차원(입력층/경쟁층)의 그리드(격자)로 구성 - 경쟁학습 실시 - 전방패스를 사용해 속도가 매우 빠름

 # 연관분석

연관분석	- 연관규칙:항목들 간의 '조건-결과'식으로 표현되는 유용한 패턴 - 이러한 패턴 규칙을 발견해내는 것을 연관분석이라고 함 - 장바구니 분석이라고 함 - 연관규칙을 찾기 위해 세분화 분석 품목이 필요하다, 다만 너무 세분화된 품목을 가지

	고 찾으려면 의미 없는 분석 결과가 도출된다.
Ariori 알고리즘	- 데이터들에 대한 발생 빈도를 기반으로 각 데이터 간의 연관관계를 밝히는 방법
FP Growth	- Apriori 단점을 보완하기 위해
장점	- 조건반응(if-then)으로 표현되는 연관 분석의 결과를 이해하기 쉬움 - 강력한 비목적성 분석 기법이며, 분석 계산이 간편함
단점	- 분석 품목 수가 증가하면 분석 계산이 기하급수적으로 증가함

 # 연관규칙 측정지표

규칙표기 : A→B

- if A then B → A가 팔리면 B가 같이 팔린다.

지지도	- 전체 거래 중 차지하는 비율을 통해 해당 연관 규칙이 얼마나 의미가 있는 것인지 확인함 - 전체 거래항목 중 상품A와 상품 B를 동시에 포함하여 거래하는 비율 - **지지도 = P(A ∩ B)** : A와 B가 동시에 포함된 거래 수 / 전체 거래 수
신뢰도	- 상품 A를 구매했을 때 상품 B를 구매할 확률이 어느 정도 되는지를 확인 - 상품 A를 포함하는 거래 중 A와 B가 동시에 거래되는 비율 - **신뢰도 = P(A ∩ B) / P(A)** : A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수
향상도 (~대비)	- A가 주어지지 않았을 때 B의 확률 대비 A가 주어졌을 때 B의 확률 증가 비율 - **향상도 = P(A ∩ B) / (P(A)*P(B))** : A와 B가 동시에 일어난 확률 / A,B가 독립된 사건일 때 A,B가 동시에 일어날 확률

 # 향상도 해석

- 향상도가 1보다 높아질수록 연고나성이 높다
- 향상도가 1보다 크면 이 규칙은 결과를 예측하는데 있어 우수하다는 것을 의미함
- 향상도가 1보다 크면 서로 양의 관계로 품목 B를 구매 할 확률보다 품목A를 구매한 후에 품목 B를 구매할 확률이 더 높다는 것을 의미함
- **향상도=1** : 품목 A와 B사이에 아무런 **상호관계없음**
- 향상도가 1보다 작으면 두 품목이 서로 음의 상관관계가 있음을 의미함

ex) 어떤 슈퍼마켓 고객 6명의 장바구니별 구입품목이 다음과 같다고 하자, 연관 규칙(콜라→맥주)의 지지도는?

거래번호	판매상품
1	소주,콜라,맥주
2	소주,콜라,와인
3	소주,주스
4	콜라,맥주
5	소주,콜라,맥주,와인
6	주스

지지도 : A와 B가 동시에 포함된 거래 수/전체 거래 수
콜라+맥주 (3) / 전체거래 (6)
3/6 = 0.5

ex) 어느 마트에서 A제품과 B제품을 판매하고 있다. A제품 → B제품의 지지도는 0.3이고, 신뢰도가 0.6이다. A제품과 B제품의 판매 수량이 동일할 때, 향상도는?

향상도 : P(A ∩ B) / (P(A)*P(B))

지지도 : P(A ∩ B) = 0.3

신뢰도 : P(A ∩ B) / P(A) = 0.6
0.3 / P(A) = 0.6
0.3 = 0.6 * P(A)
P(A) = 0.5 = 동일하다고 했으니 P(B) = 0.5

향상도 : P(A ∩ B) / (P(A)*P(B))
0.3 / 0.5 x 0.5
= 1.2

```
model = glm(default ~ ., data=Dafault, Family=binomial)
summary(model)

Call:
glm(formula = dafault ~ ., family = binomial , data = Dafault)

Deviance Residuals:
     Mln        1Q      Median          3Q           Max
-2.4691    -0.1418    -0.0557      -0.0203      3.7383

Coefficients:
                    Estimate Std.  Error    z value    Pr(>:z:)
(Intercept)       -1.087e+01   4.923e-01   -22.080     < 2e-16 ***
studentYes      -6.468e-01   2.363e-01   -2.738       0.00619   **
balance           5.737e-03 2.319e-04    24.738      < 2e-16 ***
income            3.033e-06   8.203e-06    0.370       0.71152
- - -
Signif.  codes:
0 '***' 0.001'**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

     Null deviance: 2920.6  on 9999 degrees of freedom
Residual deviance: 1571.5  on 9996 degrees of freedom
AIC: 1579.5
```