빅데이터 분석 2급 실습 보고서

가천대학교 컴퓨터공학과 201032306 유현지

 # Excel : 파일 – 옵션 – 추가기능 – 관리(이동) – 분석도구,분석도구(VBA) check!

 # Google > Kess 검색.
(Korean Educational Statics Software)

→ 통계교육 및 실무에서 필요로 하는 자료분석 기능 제공.
(Excel 버전에 맞게 download)

 # 데이터분석은 '변수'다.

- 변수가 많아질수록 조합할 수 있는 경우의 수가 많아진다.
(변수가 7개 있다면, **7!** 의 경우의 수가 나온다.)

Excel Function.

- 블록 + 더블클릭 시, 끝까지 데이터 들어감.
- choose() : Index 따라서 값 나옴.
- dataset : sheet에 담음.
- 왼쪽 모서리 박스 : 블록 잡아서 Table name 설정.
- **vlookup()**
  ex) vlookup(E2, mdindex, 2, 0)
  = mdindex 값 찾아서, mdindex table의 두번째 열에 있는 데이터 찾아옴.
- iferror(value, " ") : value값이 error이면, " "(빈칸) 설정.

Excel 실습

- CSV : 데이터가 ' **,(쉼표)**'로 구분된 파일.
- Ctrl + shift + → + ↓ = 데이터가 있는 블록 모두 잡기.
- Vlookup(): 테이블 가져와서 데이터 가져오는 함수.

** 피벗 테이블 (데이터 요약표)
- 삽입 – 피벗테이블 – 새 워크시트
- 피벗 그룹 : 시작 -15 ~ 끝 30 // 단위 5 → 5단위씩 그룹핑.
- 차트 데이터 클릭해서 '추세선 추가'
- R^2 = 0.0475 이면, 100개 중 4개만 맞춤.
  → 예측력 떨어짐. (작으면, 상관관계 작다.)
  → total 분석 시, '온도'는 제외시킬 수 있다.

- 변수 1개로는 평균이나 분산 등이 궁금. (1개로도 볼 수 있는 정보 다양) ⇒ **기술통계.**
- 2개로 비교 분석하는 것 ⇒ **상관분석 (서로 간의 연관관계).**
- 추가기능 – 통계분석 – 기술통계

- Ex) 1 2 3 3 3 3 4 4 100 → 뜬금없이 '100'이 왔는데 이것을 합해서 평균치를 낼 수 없다.
  → 그래서 **중앙값, 최빈값**을 보는 것.
- [상자그림]에서 중앙값은 '1266'인데, 2247(=보통 이상점)이 있다.

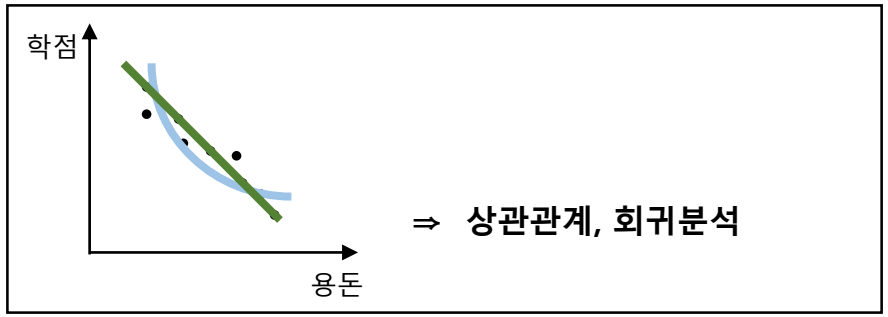
- 왜도: +일수록 왼쪽으로 치우쳐 있고, -일수록 오른쪽으로 치우쳐 있다.
  ⇒ '히스토그램'으로 시각적으로 확인 가능. (뾰족할수록 왜도값 '3'에 가깝다.)

- Total에서 평균 1290 ± 표준편차 234 가 (평균의 표준편차 7.08) 70%의 data가 있다.
- 줄기잎그림 '세로 ' 로 돌리면, 히스토그램 된다.

Excel 실습

- 표준편차 클수록 변동이 심하고, 작을수록 평준화 되어있다.
- 평균분석은 집단끼리 비교할 때, 많이 쓰임.
- 변동계수 클수록 변동이 심하다.

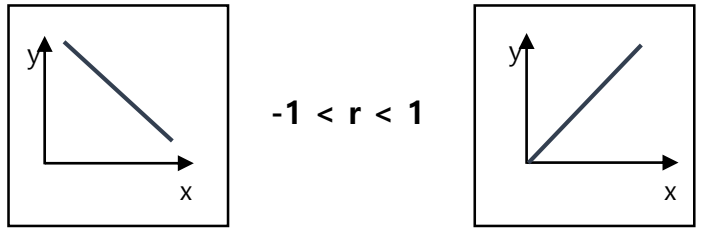
- 산점도 : 점 하나가 '객체 '



(산점도 그릴 땐, 숫자데이터를 X축 변수: temp / Y축 변수: smoothie 설정 후, 그래프 – '추세선추가'로 R^2 값 확인.)

- 상관분석 : X값이 늘어갈 때, Y도 변할까?

- 쌍대(?)비교 해준다.
  (숫자 있는 것 모두 변수로 넣어준다.)



음의 관계

-1 ~ -0.6

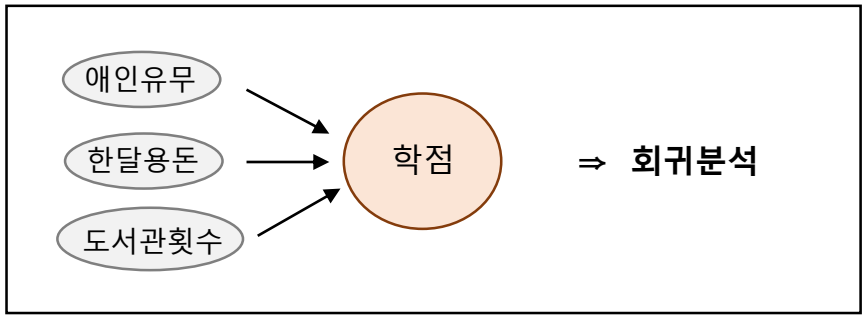
양의 관계

0.6 ~ 1

영향력 있는 구간.

Excel 실습

- 회귀분석 : 방정식을 만들어주고, 결정계수는 그 확률!! ⇒ 머신러닝!!
  (R^2 이 결국 결정계수) : 산점도에서 추세선 그려보면, 알 수 있다.



(이러한 정보들을 알면, 학점을 예측할 수 있다.)

- 회귀분석을 하기 위해선, 상관관계를 먼저 알아야함.
- 상관관계(상관계수) 높은 것만 가지고, 회귀분석!!

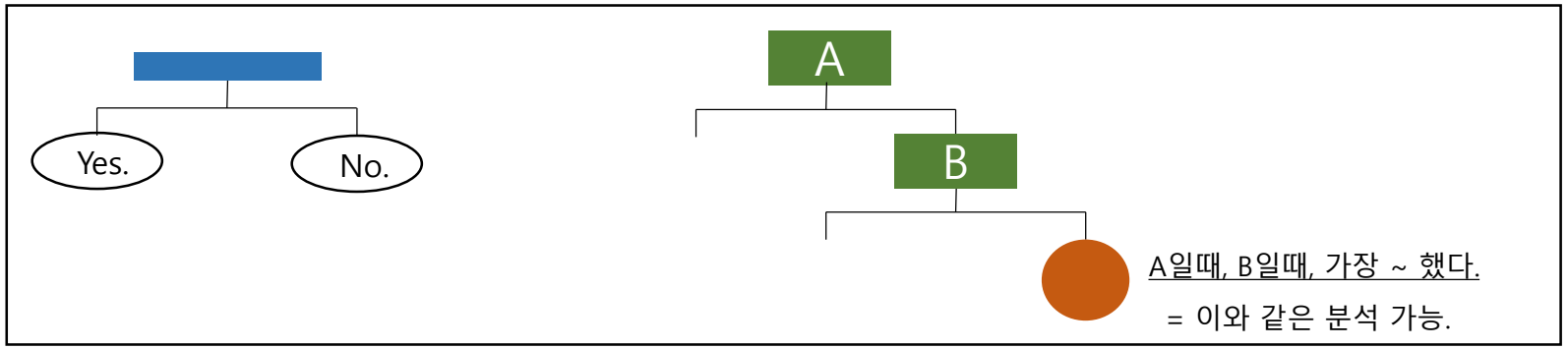
- 독립변수 1개 일 때, $y = ax + b$
- 독립변수 2개 일 때, $y = a_1x_1 + a_2x_2 + b$
  ⇒ 선형 방정식

- 결정계수가 0.9 이면, 확률 높다.

Excel 실습

- 다변량 분석(변수 여러 개) → 의사결정나무 (의사결정트리)

- 반응변수: 질산의 양 (연속형 변수: 숫자 / 명목형 변수: A, B ..)
- 분류표, 분류결과 모두 표시.



- Ex) 교통사고 가장 많은 구간은
  ~~~~~한 곳이며,
  ~~~~~이며,
  ~~~~~이며,
  ~~~~~인 곳입니다.

⇒ 이러한 결과 도출 가능.

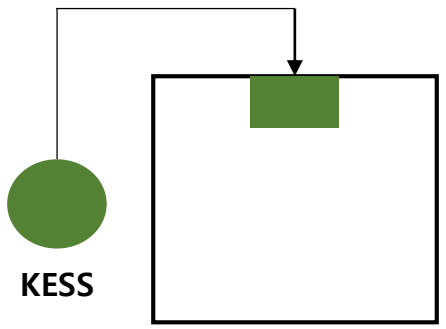
Excel 실습

- 군집분석
  - 계층형 (가까운 데이터끼리 묶는 것.)
  - K-means (직관적)

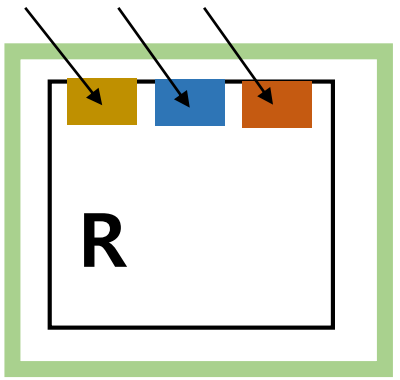
- 라벨변수: id
- 분석변수: ~
- 군집갯수: 4개 / 최종군집기록 check!

Ex) 키가 크지만 허리둘레, 엉덩이둘레가 작은 그룹....등
이렇게 그룹별로 나누어 분석.

- 1번 군집에 얼만큼의 데이터가 있는지는 '군집크기'로 나옴.



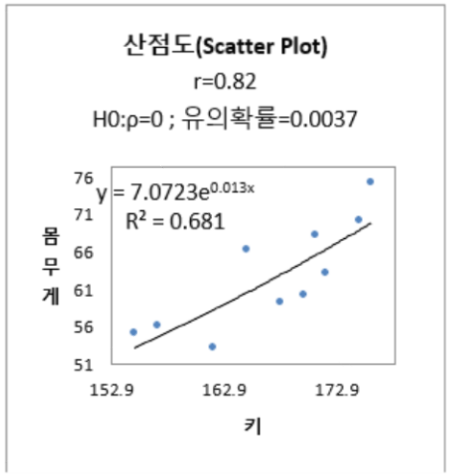
KESS 다운받아서 엑셀에 설치.



상관분석

상관계수
(유의확률)

	키	몸무게
키	1	0.8192
(유의확률)	.	0.0037
몸무게	0.8192	1
(유의확률)	0.0037	.



회귀분석결과

분산분석표

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	310.3908	1	310.3908	16.325	0.0037
잔차	152.1092	8	19.0136		
계	462.5000	9			

Root MSE	4.3605
결정계수	0.6711
수정결정계수	3.9600

모수 추정

몸무게 = -73.45361 + 0.81361*키

변수명	추정값	표준오차	t-통계량	유의확률
절편	-73.45361	33.67700	-2.181	0.0608
키	0.81361	0.20137	4.040	0.0037

회귀분석결과

분산분석표

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	1880.4428	2	940.2214	89.642	< 0.0001
잔차	188.7953	18	10.4886		
계	2069.2381	20			

Root MSE	3.2386
결정계수	0.9088
수정결정계수	2.8248

모수 추정

변수명	추정값	표준오차	t-통계량	유의확률
절편	-50.35884	5.13833	-9.801	< 0.0001
물의온도	1.29535	0.36749	3.525	0.0024
공기주입량	0.67115	0.12669	5.298	< 0.0001

의 양 = (1.29535*물의온도) + (0.67115*공기주입량) + -50.3

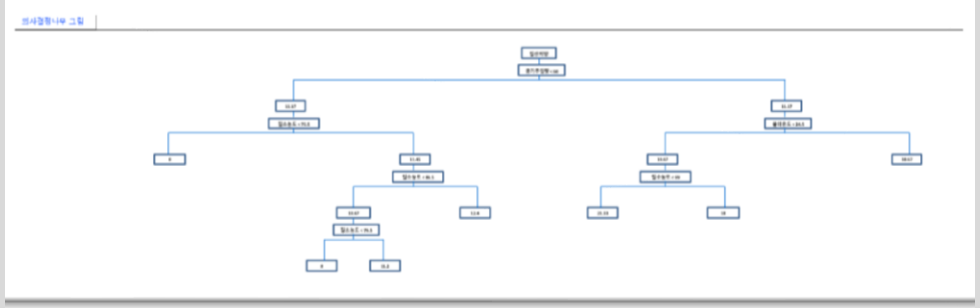
**** 선형회귀분석 ****

상관분석결과

상관분석

상관계수
(유의확률)

	공기주입량	물의온도	질소농도	질산의양
공기주입량	1	0.7819	0.5001	0.9197
(유의확률)	.	0	0.0209	0
물의온도	0.7819	1	0.3909	0.8755
(유의확률)	0	.	0.0797	0
질소농도	0.5001	0.3909	1	0.3998
(유의확률)	0.0209	0.0797	.	0.0725
질산의양	0.9197	0.8755	0.3998	1
(유의확률)	0	0	0.0725	.



R-실습

```r
a <- "홍길동"  # 스칼라 변수 생성
av <- c("홍길동","김철수","김영희","김순이")  # 이름 벡터 생성
bv <- c(23, 34, 45, 32)  # 나이 벡터 생성
cv <- c(3.4, 1.5, 4.2, 3.9)  # 벡터 학점 생성
dv <- c("서울","인천","수원","성남")  # 주소 벡터 생성

edf <- data.frame(av, bv, cv, dv)  # 데이터프레임 생성
names(edf) # 데이터프레임의 벡터 이름 가져오기
names(edf) <- c("이름", "나이", "학점", "주소")  # 데이터프레임 속 벡터이름 설정

edf$이름  # 이름만 보고싶을 때
edf$나이  # 나이만 보고싶을 때

edf$주소[2]  # edf 데이터프레임의 주소벡터에서 두번째값 가져올때
edf[2]  # edf 데이터프레임에서 두번째 값 = 두번째 벡터
edf[2,3]  # edf 에서 2행 3열
edf[ ,3]  # 모든 행을 보여주되, 3행만 보여줌.
edf[2, ]  # 2행이 보여주고, 열은 모두 보여줌.
```

```r
## google비즈라는 패키지
install.packages("googleVis")
library(googleVis)

# 해당시간에 따라 몇명이 나갔고 들어왔는지 보여주는 데이터.
loc <- read.csv("R/14subway.csv",header=T)
head(loc)

# gvisMotionChart() 함수로 시각화 그래프 그려보기.
t1 <- gvisMotionChart(loc, idvar="line_no", timevar="time",
                      options=list(width=1000,height=500))
plot(t1)
```

```r
# 구글지도 API
install.packages("ggmap")
library(ggmap)

install.packages("ggplot2")
library(ggplot2)


# cctv data 가져와서 저장.
cctv <- read.csv("R/koreacctv.csv", header = T)
head(cctv)
# 데이터프레임으로 만들어줌.
cctv <- as.data.frame(cctv)
str(cctv)
# cctv 데이터프레임에서 3,4,11,12열만 가져오기.
cctv2 <- cctv[  , c(3,4,11,12)]
head(cctv2)


# 경도와 위도를 cent에 담는다.
cent <- c(lon=127, lat=37.6)

# ggmap 패키지에서 get_googlemap()함수를 사용.
# 위에서 잡은 center를 지도의 중심으로 잡는다.
# 전체 세계지도 zoom=1, 숫자클수록 확대.
map2 <- ggmap(get_googlemap(center = cent, zoom=13, maptype='roadmap', color='bw'))
map2 # 지도 실행.
names(cctv)[11] <- "lat"
names(cctv)[12] <- "lon"
map2 + geom_point(data=cctv, aes(x=lon, y=lat), colour = 'gray10', alpha=0.6)
```

THANK YOU