데이터과학을 위한 R프로그래밍

5주차. 데이터탐색

이혜선 교수
포항공과대학교 산업경영공학과

목차

5주차. 데이터탐색

5주차

2차시

데이터탐색과 기술통계치

POSTECH

데이터 기술통계치요약

- 데이터 : 학생들의 학업성취도* (포루투칼의 고등학생 수학점수)
- http://archive.ics.uci.edu/ml/datasets/Student+Performance

stud_math.csv

	A	B	C	D	E	F	G	H	I	J	K	L	
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	trav
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	
5	GP	F	15	U	GT3	T	4	2	health	services	home	mother	
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	
7	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	
8	GP	M	16	U	LE3	T	2	2	other	other	home	mother	
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	
10	GP	M	15	U	LE3	A	3	2	services	other	home	mother	
11	GP	M	15	U	GT3	T	3	4	other	other	home	mother	
12	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	
13	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	
14	GP	M	15	U	LE3	T	4	4	health	services	course	father	
15	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

데이터탐색과 기술통계치

POSTECH

데이터 기술통계치요약

데이터설명 (stud_math_desc.doc참고)

school : 학교이름 (GP, MS)

sex : 성별 (F, M)

age : 나이 (15-22)

address : 주소 (Urban:도심, Rural:외곽)

Medu : 엄마교육수준

famsize : 가족수 (LE3 :≤3, GT3: >3)

Fedu : 아빠교육수준

Traveltime : 통학시간 1(15분이하),2,3,4(1시간이상)

Dalc : 음주(1-5)

Studytime : 주중공부시간: 1(<2시간), 2(2-5시간), 3(5-10시간), 4(>10시간)

health : 건강상태 (1(매우나쁨)-5(매우좋음))

activities : 방과후활동(yes, no)

romantic : 이성교제여부(yes, no)

Nursery : 유치원다녔는지여부(yes, no)

soout : 친구들과 외출 (1-5)

internet : 집에서 인터넷사용(yes, no)

absences : 학교결석 (0-93)

타겟변수 : G3(최종성적, 0-20), G2(2학년), G1(1학년)

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary edu
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary educa
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

POSTECH

데이터 기술통계치요약

- 데이터 : 학생들의 학업성취도* (포루투칼의 고등학생 수학성적)
  - stud 데이터는 n=395관측치와 33개의 변수

```
# Data exploration : Numerical summary statistics
library(dplyr)

# set working directory
setwd("D:/tempstore/moocr")

### student math grade data ####

stud<-read.csv("stud_math.csv")

head(stud)
dim(stud)
str(stud)

attach(stud)
```

```
> dim(stud)
[1] 395  33
> str(stud)
'data.frame':   395 obs. of  33 variables:
 $ school     : chr  "GP" "GP" "GP" "GP" ...
 $ sex        : chr  "F" "F" "F" "F" ...
 $ age        : int  18 17 15 15 16 16 16 17 15
 $ address    : chr  "U" "U" "U" "U" ...
 $ famsize    : chr  "GT3" "GT3" "LE3" "GT3" ..
 $ Pstatus    : chr  "A" "T" "T" "T" ...
 $ Medu       : int  4 1 1 4 3 4 2 4 3 3 ...
```

POSTECH

데이터 기술통계치요약

```
# set working directory
setwd("D:/tempstore/moocr")

### student math grade data ####
stud<-read.csv("stud_math.csv")
str(stud)

# character variable to factor
stud<-read.csv("stud_math.csv",stringsAsFactors = TRUE)
str(stud)
```

```
> stud<-read.csv("stud_math.csv")
> str(stud)
'data.frame':   395 obs. of  33 variables:
 $ school     : chr  "GP" "GP" "GP" "GP" ...
 $ sex        : chr  "F" "F" "F" "F" ...
 $ age        : int  18 17 15 15 16 16 16 17 15 15
 $ address    : chr  "U" "U" "U" "U" ...
 $ famsize    : chr  "GT3" "GT3" "LE3" "GT3" ...
 $ Pstatus    : chr  "A" "T" "T" "T" ...
 $ Medu       : int  4 1 1 4 3 4 2 4 3 3 ...
```

```
> stud<-read.csv("stud_math.csv",stringsAsFactors = TRUE)
> str(stud)
'data.frame':   395 obs. of  33 variables:
 $ school     : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1
 1 ...
 $ sex        : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2
```

POSTECH

데이터 기술통계치요약

- summary(데이터이름) : 각 변수별로 요약통계량을 제공.
  - 숫자변수에 대해서는 (최소값, 25%, 중위값, 평균, 75%, 최대값)을 제공

```
# summary statistics for numerical variables
summary(stud)
```

```
> # descriptive statistics
> summary(stud)
 school    sex          age          address
 GP:349    F:208   Min.   :15.0    R: 88
 MS: 46    M:187   1st Qu.:16.0    U:307
                   Median :17.0
                   Mean   :16.7
                   3rd Qu.:18.0
                   Max.   :22.0
```

POSTECH

데이터 기술통계치요약

- mean(변수) : 평균
- sd(변수) : 표준편차 (분산의 제곱근)
- var(변수) : 분산

```
mean(G3)
sd(G3)
sqrt(var(G3))
```

```
> mean(G3)
[1] 10.41519
> sd(G3)
[1] 4.581443
> sqrt(var(G3))
[1] 4.581443
```

통계함수	설명
mean(x)	평균
median(x)	중앙값
sd(x)	표준편차
mad(x)	Median absolute deviation
var(x)	분산

POSTECH

데이터 기술통계치요약

- 특정변수들에 대한 요약통계량 (dplyr 활용 - lec3_3.r)
- : select(stud, c("변수1", "변수2", "변수3"))
  %>% summarize_all(FUN)
  - stud데이터는 33개의 변수를 가짐!! ⇒ 특정변수들에 대해 탐색하고자 할 때

```
# summarize with interested variable list using dplyr(lec3_3.r)
a1 <- select(stud, c("G1", "G2", "G3")) %>% summarize_all(mean)
a2 <- select(stud, c("G1", "G2", "G3")) %>% summarize_all(sd)
a3 <- select(stud, c("G1", "G2", "G3")) %>% summarize_all(min)
a4 <- select(stud, c("G1", "G2", "G3")) %>% summarize_all(max)
table1 <- rbind(a1,a2,a3,a4)
rownames(table1) <- c("mean","sd","min","max")
table1
```

```
> table1 <- rbind(a1,a2,a3,a4)
> rownames(table1) <- c("mean","sd","min","max")
> table1
             G1        G2        G3
mean 10.908861 10.713924 10.415190
sd    3.319195  3.761505  4.581443
min   3.000000  0.000000  0.000000
max  19.000000 19.000000 20.000000
```

POSTECH

데이터 기술통계치요약

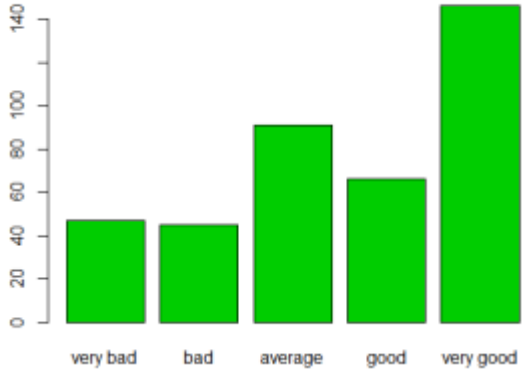
범주형 변수의 요약 : table(변수이름)

```
# categorical data
table(health)
```

```
> table(health)
health
  1   2   3   4   5
 47  45  91  66 146
```

막대그림 (이름주기)

```
health_freq<-table(health)
names(health_freq) <- c ("very bad", "bad", "average", "good",
                         "very good")
barplot(health_freq, col=3)
```



POSTECH

데이터 기술통계치요약

범주형 변수의 요약 : table(변수1, 변수2)

2*2 분할표

```
# 2*2 contingency table
table(health,studytime)
```

```
> table(health,studytime)
      studytime
health  1  2  3  4
     1 11 29  3  4
     2  9 27  6  3
     3 20 43 18 10
     4 15 30 17  4
     5 50 69 21  6
```