Midterm test

Data Science – 2020 Spring Semester

2020-04-24

7:00 ~ 9:00 PM

"**midterm2020.RData**" contains the datasets for this midterm test. After loading this file into R, perform following tasks. You need to submit **a reporting document** that describe all your answers and the process and explanation how you derive the answers (MS word, pdf, hwp, ppt, or any of your preferred form), **R script file** that contains the r code you use for this test, and **RData file** that has result data.frames from the questions. All the resulting data frames should be in **one RData file** named **st#######.RData** where ###### represents your student ID. You can choose to answer in your proficient language from English and Korean. The **submission should be done before 9 PM** otherwise the submission page will be locked and late submission will not be counted for grading. You can submit your work several times within the time limit.

"**midterm2020.RData**" 파일은 이번 중간고사에서 사용하게 될 dataset을 저장하고 있는 RData 파일입니다. 이 파일을 R에 loading한 후 아래 task를 수행하시오. task를 수행하면서 각 질문의 **답과 답을 도출하게 된 과정 및 설명을 보고서**(문서, word, pdf, 한글, ppt 등)에 작성하여 edmodo에 제출하시오. 문제를 풀기 위해 사용한 **r script 파일**과 결과 **RData 파일**(하나의 RData 파일에 다 넣어서)도 함께 제출하시오. 문제의 답으로 나오는 결과 data frame은 여러 개인데, 모두 하나의 RData 파일에 넣어서 제출하고 파일의 이름은 **st#########.RData** 입니다. #######은 본인 학번, 형식이 틀리는 경우 감점이 있을 수 있습니다. **9시가 지나면 제출이 불가능**하므로 시간관리를 잘해주세요. 답은 영어나 한글 중 편한 언어를 쓰면 됩니다. 늦게 제출한 부분은 성적에 반영되지 않음을 기억하세요. 9시 전이라면 몇 번이라도 반복해서 제출 가능합니다.

Data Exploration and Preparation

"pums.sample" data frame contains sample dataset from the United States Census 2011 national

PUMS American Community Survey. The sample dataset is about mostly Americans who has a job. Attributes are given as follows.

(KOR) **"pums.sample"** data frame은 2011년 미국 인구조사 데이터의 일부 sample을 저장하고 있는 데이터이다. 이 중에서 미국에서 직업을 가지고 일을 하고 있는 사람을 대상으로 조사한 여러 정보를 담고 있다. 각 변수는 다음과 같은 의미를 지닌다.

PINCP: person's yearly income (in USD)

SEX: Sex: 1 for male, 2 for female

AGEP: person's age

MAR: martial state

COW: class of work(job)

SCHL: education level

WKHP: average weekly working hours

FER: whether giving birth to child for last 1 year (1년동안 자녀를 출산한 적이 있는지)

JWMNP: commuting time in minute (직장까지의 통근 시간 (단위 분))

DIS: disability of person (장애가 있는 지 여부)

1. Describe the structure of dataset. How many observations and variables are there? What are the types of the variables? Answer if each variables is properly typed with explanation.

   (KOR) 데이터의 구조를 파악하시오. 데이터의 observation의 수와 variable의 수는 어떻게 되나요? 각 variable의 type은 어떠한가요? 각 variable의 type이 적절하게 설정되어 있는지 그렇지 않은지 설명하시오.

2. For the variable **SEX**, value 1 is male person and 2 is for female person. To clarify the meaning of data, change the value 1 and 2 into "**Male"** and "**Female"**, respectively. Set the variable with its proper type with your explanation.

   (KOR) 변수 "**SEX**"는 조사 대상자의 성별을 의미한다. 1은 남성, 2는 여성을 의미한다. 의미를 명확히 하게 위해 1과 2대신 Male, Female로 값을 변경하시오. 변수의 type은 적절한 type이 무엇인지 생각하여 해당 type으로 변경하시오. 왜 그 type으로 설정했는지 설명하시오.

3. Variable **MAR** means marital status for person. It has value from 1 to 5. To clarify the meaning, change the values as follows and type of variables if necessary.

   (KOR) 변수 "**MAR**"는 marital status를 의미합니다. 1부터 5까지의 숫자로 표시되어 있는데, 각각의 의미를 쉽게 파악할 수 있도록 값을 변경하고 type또한 변환하시오.

   - ✓ 1 .Married

- ✓ 2 .Widowed
- ✓ 3 .Divorced
- ✓ 4 .Separated
- ✓ 5 .Never married or under 15 years old

4. Examine the missing value of variables. What are the variables with no missing values? What are the variables that has missing values and what is the ratio of missing values for the variables?

   (KOR) 각 변수의 Missing Value의 수는 어떻게 되는가? Missing Value가 없는 변수는 어떤 변수이며, Missing Value가 있는 변수는 전체의 몇 %가 Missing 되어 있는지 비율을 계산하시오.

5. Variable **FER** means maternity for last 12 months. If the person is male or female younger than 15 or older than 50, the value is missing. Examine all male people have missing value for the variable. For female person, examine whether the values are missing for all the women who are younger than 15 and older than 50. State if there are any missing values under any other condition stated here. How many are there if any?

   (KOR) 변수 "**FER**"은 과거 12개월 동안 출산여부를 나타낸다. 이 변수의 Missing value는 남성이거나 나이가 15세 미만 또는 50세 초과인 경우 나타난다고 한다. 모든 남성에 대해 FER이 Missing인지 확인하시오. 또한 여성의 경우 15세 미만인 경우 모두 Missing인지 확인하시오. 또 50세 초과 여성이 모두 Missing인지 확인하시오. 여기서 언급된 경우 외에 Missing Value가 발생했는지 했다면 얼마나 발생했는지 확인하시오.

6. Examine if there are any outliers for each variable. If there are any outliers which are obvious errors, then state and explain why they are obvious errors. If they are not obvious errors and need to be confirmed, then explain why it is so.

   (KOR) 각 변수의 outlier가 존재하는지 확인하시오. 발견한 outlier 중 명백한 오류가 있는지 확인하고, 왜 그 outlier가 명백한 오류인지(혹은 명백한 오류가 아닌지) 설명하시오.

7. Compare the variable SCHL and COW. Can you say that person's class of work is different according to his level of education? Examine that from sample data and explain your findings.

   (KOR) 교육수준(SCHL)과 종사업종(COW)를 비교해보시오. 교육수준에 따라 종사업종의 종류나 비율이 어떻게 달라지는지 (혹은 달라지지 않는지) 확인하고 설명하시오.

8. Partition the sample into 5 age groups of 20s, 30s, 40s, 50s, and over 60. (include persons into group of 20s if he/she is under age). Compare the yearly income **PINCP** with each age

group. What are the groups with the highest and lowest average yearly income?

(KOR) 나이대별 소득을 비교해보시오. 20대, 30대, 40대, 50대, 60대 이상, 5개의 그룹으로 나누어 각 나이 그룹별 평균 소득을 비교해보시오. (20세보다 어린 사람이 있다면 20대에 추가하세요.) 가장 소득이 높은 그룹과 가장 소득이 낮은 그룹은 어디인가요?

9. Try to draw some charts to compare WKHP and PINCP variable. How the variables are related? Explain the relation between two variables that you found from the chart.

   (KOR) 근무시간(WKHP)과 소득(PINCP)과 어떤 관계가 있는지 그래프를 그려 확인해보시오. 그래프에 관계가 나타난다면 그것을 설명해보시오.

10. Save the result data.frame in a RData file (with save command) and submit the file to Edmodo.

    (KOR) 결과 data.frame을 RData 파일로 저장해서 함께 제출하시오. (save 명령 사용)

Data Transformation

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
...
50           5.0         3.3          1.4         0.2     setosa
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
...
100          5.7         2.8          4.1         1.3 versicolor
101          6.3         3.3          6.0         2.5  virginica
102          5.8         2.7          5.1         1.9  virginica
103          7.1         3.0          5.9         2.1  virginica
...
150          5.9         3.0          5.1         1.8  virginica
```

The iris dataset is built in R. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

Transform the iris data into following forms. (iris dataset을 아래와 같이 변환하시오)

1. iris.wide

```
> iris.wide
      Species  Part Length Width
1      setosa Petal    1.4   0.2
2      setosa Sepal    5.1   3.5
3      setosa Petal    1.4   0.2
4      setosa Sepal    4.9   3.0
5      setosa Petal    1.3   0.2
6      setosa Sepal    4.7   3.2
7      setosa Petal    1.5   0.2
8      setosa Sepal    4.6   3.1
9      setosa Petal    1.4   0.2
10     setosa Sepal    5.0   3.6
11     setosa Petal    1.7   0.4
12     setosa Sepal    5.4   3.9
```

```
> str(iris.wide)
'data.frame':300 obs. of  4 variables:
 $ Species: Factor w/ 3 levels "Setosa",..: 1 1 1 1 1 1 1 1 1 ...
 $ Part   : chr  "Petal" "Petal" "Petal" "Petal" ...
 $ Length : num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Width  : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
```

2. iris.tidy

```
> str(iris.tidy)
'data.frame':600 obs. of  4 variables:
 $ Species: Factor w/ 3 levels "Setosa",..: 1 1 1 1 1 1 1 1 1 ...
 $ Part   : chr  "Sepal" "Sepal" "Sepal" "Sepal" ...
 $ Measure: chr  "Length" "Length" "Length" "Length" ...
 $ Value  : num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
```

```
> iris.tidy
      Species  Part Measure Value
1      setosa Sepal  Length   5.1
2      setosa Sepal  Length   4.9
3      setosa Sepal  Length   4.7
4      setosa Sepal  Length   4.6
5      setosa Sepal  Length   5.0
6      setosa Sepal  Length   5.4
7      setosa Sepal  Length   4.6
8      setosa Sepal  Length   5.0
9      setosa Sepal  Length   4.4
10     setosa Sepal  Length   4.9
11     setosa Sepal  Length   5.4
12     setosa Sepal  Length   4.8
13     setosa Sepal  Length   4.8
14     setosa Sepal  Length   4.3
15     setosa Sepal  Length   5.8
```

include **iris.wide** and **iris.tidy** in the RData file (st########.RData) and submit the file.

iris.wide와 **iris.tidy** 모두 RData 파일 (st########.RData) 에 저장해서 제출하세요.