Some R Problems Derived Me Nuts!

iHELP Working Group

April 23, 2013

Chapter 1

서문

R을 사용하게 된 이래로 여러가지 경험들을 바탕으로 도움이 되고자 하는 내용을 정리하는 페이지입니다.

본 페이지는 다양한 종류의 패키지들을 활용하는 방법보다는 R 기본 시스템에서 제공되어진 기능들만을 이용하여 원리를

이 문서는 매일 갱신이 되며, 문서의 최초 작성일은 2013년 4월 10일입니다.

만약, 이 페이지를 읽고 있는 사용자가 이 문서가 도움이 되었다고 생각된다면 문서에 더 많은 내용이 기재될 수 있도록 추가, 수정 및 제안에 대한 내용을 ihelp-urquestion@lists.r-forge.r-project.org의 주소로 이메일을 보내주시길 부탁드립니다.

본 페이지는 iHELP Working Group의 관리자에 의해서 정리가 되고 있지만, 무수한 ihelp-urquestion 메일링 리스

메일링에 등록된 내용은 이 문서의 관리자가 문서를 갱신할 때마다 반영하도록 노력할 것입니다.

1.1 미분류 질문들 (이 섹션에 등록된 질문들은 접수만 되고 아직은 답변되지 않은 상태입니다)

1. (제안) 본 페이지 자체를 데이터 분석을 위한 프로세스를 나타낼 수 있도록 섹션구성 자체를
   - 최초 데이터의 입력과 확인
   - 데이터를 조작
   - 데이터 정리하기
   - 통계로 데이터를 설명하기
   - 그래픽적으로 보여주기
   - 원하는 그래픽 만들기

   으로 조정함. – 즉각적 수용 – 문서구조 변경함 (2013-APR-24)
2. 데이터의 형변환과 결측치에대 대한 전반적인 내용을 알려주면 될 듯함.
3. 기본 데이터 형에 대해서 – 벡터, 행렬, 배열, 데이터 프레임, 리스트, 요인, 그리고 이들을 다루는데 필요한 함수들

```
ls()

names(mydata)

str(mydata)
```

```
dim(object)

class(obj)

mydata

head(mydata, n=10)

tail(mydata, n=5)

length(obj)
str(obj)
class(obj)
names(obj)

c(obj,obj,...)
cbind(obj, obj, ...)
rbind(obj, obj, ...)

obj

ls()
rm(obj)

newobject <- edit(obj)
fix(obj)
```

4. 데이터 입출력

```
mydata <- read.table("c:/mydata.csv", header=TRUE, sep=",", row.names="id")
write.table(mydata, "c:/mydata.txt", sep="\t")
```

5. R을 사용하기 전에 반드시 알아두어야 할 점 – R은 모든 연산을 열벡터를 기준으로 한다는 것을 반드시 알고 시작해야 함. SAS는 행벡터임. 따라서, 가끔 R에서 벡터사이즈가 어쩌구 할때는 바로 데이터의 수가 R이 해결하기에는 부족하다는 점이다. 그런데, 어떤 경우 이것은 주로 메모리 조절과 관계가 있음.

6. (접수: 2013-APR-23) NA 와 NaN을 데이터로부터 찾고 싶어요.

   (답변) 요것은.... `is.na()`와 `is.nan()` 함수 사용법을 알려주면 좋음. 추가로 `is.null()`도 알려주면 `is`관련 함수들에 설명해주면 짱임.

7. (접수: 2013-APR-23) 분기문 쓸때요 `if`문 사용하는 것은 알고 있습니다. 그런데, C 언어처럼 중간에 루프를 완전히 끊고 나가는 방법이 있거나 혹은 해당 루프를 넘어가는 방법이 있나요? 예를들면, Basic 언어에서 보면 goto 같은 것도 있는지 알고 싶어요.

   (답변) 요건... 질문이 좀 방대한건데... `repeat`, `while`에 대해서 간단히 보여주고, `break`과 `next` 를 알려주면 좋음. 만약, 메시지 역시 설명해줄려면 `stop()`, `warning()`도 함께 설명해주면 최고임.

8. (접수: 2013-APR-23) 저는 다각형을 그리고 싶습니다.

   (답변) 이것은 간단히 2차원-랜덤포인트 생성한 뒤에 `polygon()` 함수를 써서 보여주면 됨.

9. (접수: 2013-APR-23) 제가 가진 데이터셋이 있는데, 이 데이터를 어떤 특정한 변수들의 값을 이용하여 분류하려고 합니다. 어떻게 해야하나요?

   (답변) 요것은 `split()` 함수를 이용하도록 알려줄 것.

10. (접수: 2013-APR-23) 제가 가진 데이터 프레임에 NA 값들이 있는데, NA 때문에 분석이 이상해지는 것 같아, NA를 가진 데이터 행자체를 없애고 싶습니다. 한번에 해주는게 없나요?

    (답변) 요건 `na.omit()`과 같은 함수를 이용하는 법을 알려줄 것. 흠.. `na.action`이라는 개념을 알려주면 더욱 좋음.

11. (접수: 2013-APR-23) 논리형 벡터가 있는데, 이 벡터의 구성요소가 모두 TRUE 인지 알고 싶습니다.

    (답변) 이건 `isTRUE()`함수와 `all()`함수를 통해 알려주면 매우 좋음.

12. (접수: 2013-APR-23) t-테스트 하는 방법 좀 알려주세요 – 통계적 해석을 덧 붙여주시면 좋을 것 같습니다.

    (답변) `t.test()` 함수 사용법을 알려줄 것 – 일반화 된 옵션 다 알려주면 더 좋을 것 같음.

13. (접수: 2013-APR-22) 선형방정식 $AX = b$의 해 X를 찾으려면 어떻게 해야 하나요?

    (답변) `solve()` 함수의 사용법을 알려줄 것.

14. (접수: 2013-APR-21) R 패키지를 CRAN에 올리는 방법을 알려주세요

    (답변) 이 질문을 대답할 때는 반드시 CRAN Package Submission Guideline에 대해서 알려줘야 함. (이거 번역해 놨는데 당췌 어디에 뒀는지 찾을 수가 없음, 2013-04-20 까지 못 찾으면 새로이 번역할 것)

15. (접수: 2013-APR-19) R은 처음부터 기존의 통계팩키지와는 다른 모습에 약간 두렵기까지 합니다. 기존의 분석은 일반적으로 [프로그램 실행 – > 데이터 불러오기 – > 분석(메뉴클릭:SPSS 또는 명령어입력:SAS) – > 실행]의 절차를 밟아 왔기에 모든 결과를 한 번에 보여주는 식입니다. 그러나 R은 그렇지 않아 이러한 점부터 생소하고 이상합니다. 데이터를 불러오기 하면 바로 데이터시트를 볼 수 있는 것도 아닙니다 (접수날짜: 2013-APR-17).

    (답변) 사용방식의 다른 점에 대해서 아주 근본적인 다른 점을 알려줄 것. Introduction to R 문서에 써 있음 (단순히 링크시켜주는게 좋을 듯 함).

    파일관리와 관계된 여러가지 유용한 유틸리티가 존재합니다.

    - `edit()`,
    - `file.edit()`
    - `fix()`
    - `file.show()`
    - `file.path()`
    - `list.files()`
    - `dir.create()`
    - `file.access()`
    - `file.exists()`
    - `file.copy()`
    - `data.entry()`
    - 너무 많아서 차근차근 예를들면서 하나씩 설명하겠습니다.

16. (접수: 2013-04-18, Reproducibility=NA) `c()`의 역할은 무엇인가요? `a <- seq(1:4)`과 `a <- c(1,2,3,4)`은 동일한 것인가요?

17. (접수: 2013-04-18, Reproducibility=NO) read.xlsx함수를 이용해 xlsx파일에서 데이터프레임형태로 가져옵니다. 이 때 [3,3] 셀에 있는 텍스트가 "3월" 이라고 할 때 temp[3,3] == "3월" 이렇게 비교하려고 하면 제대로 비교가 안되더군요.. 한글 텍스트로 이루어진 변수값를 비교하는 방법이 어떤게 있는지 궁금합니다.

18. 분석을 하고 나면 결과를 그래프나 그림으로 나타내게 되는데 R에서는 그림을 나타내는 창이 하나만 나타나서 동시에 두 개를 보지 못하는 경우가 허다한데, 이의 해결방법은 없나요? (접수: 2013-APR-13, 분류: 그래픽스 관련)

    (답변) R에서는 그래픽 디바이스가 그래픽 생성시 마다 초기화되어 다시 보여줌으로서 그래픽 창이 하나만 계속 보여지는 것입니다. 새로운 그래프를 또다른 장치를 통해 보여주고자 한다면 `X11()`이라는 명령어를 이용하면 됩니다. 이 명령어는 유닉스환경에 설치된 R의 경우에 해당합니다.

19. 초기에 가장 보는 에러는 "xxx 함수가 없습니다" 또는 "xxx 함수를 찾을 수 없습니다"입니다. (접수: 2013-APR-15, 분류: 패키지 관련)

    (답변) 대부분의 경우는 사용하고자 하는 함수가 R 기본 배포판에 포함되어 있지 않은 사용자에 의해서 제공된 특정한 패키지에서 존재하기 때문입니다. 이런 경우에는 먼저 사용하고자 하는 함수가 어떤 패키지에 존재하는지 알아야 합니다. 그리고, 해당 패키지를 설치했을 때에는 설치된 패키지를 사용할 수 있도록 로딩하는 과정을 거쳐야 합니다.

    ```
    > library(pkg_name)
    ```

    이와 반대로 현재 연결된 라이브러리를 떼어낼 수도 있습니다.

    ```
    > detach(package:pkg_name)
    ```

20. 패키지를 설치 (분류: 사용자 환경)

    (답변) 설치되는 패키지의 **설치위치**와 **의존성**에 대해서 반드시 알아야 합니다.

    ```
    > install.packages("패키지명", dependencies=TRUE, )
    ```

21. 설치된 패키지의 목록을 확인하는 방법을 알고 싶습니다.

    (답변)

22. (접수: 2013-APR-15) `setwd()` 와 `getwd()`를 활용하기

    (답변)

23. (접수: 2013-APR-15)

    (답변)

1.2 프로그래밍 언어를 처음 접하거나 전산지식이 전무하신 분들을 위하여

1. 프롬프트가 무엇인가요?

    ```
    >
    ```

    (답변)

1.3 데이터 조작에 관련하여

분석자가 보통 얻게 되는 데이터는 분석에 사용되는 통계모형에 적합한 경우는 드물기 때문에 분석자 스스로가 이러한 데이터를 형성하는 것은 필요한 기술중에 하나라고 할 수 있습니다.

1. 아래와 같이 주어진 데이터에 변수 ID는 결측값 없이 모든 값이 완전하게 잘 들어가 있는데, Week 변수에는 각 ID의 첫번째 레코드에만 해당하는 부분에 값이 들어가 있고 나머지부분에는 `NA`값이 들어가 있습니다.

```
mydata <- data.frame(ID=c(rep(1,4), rep(2,4), rep(3,2)), Week=c(15, NA, NA, NA, 18, NA, NA, NA, 20,
```

```
> mydata

   ID Week
1   1   15
2   1   NA
3   1   NA
4   1   NA
5   2   18
6   2   NA
7   2   NA
8   2   NA
9   3   20
10  3   NA
```

이와 같은 데이터를 아래와 같이 자동으로 채워주려면 어떻게 해야 할까요?

```
   ID Week
1   1   15
2   1   15
3   1   15
4   1   15
5   2   18
6   2   18
7   2   18
8   2   18
9   3   20
10  3   20
```

이를 수행하는데에는 여러 가지 종류의 함수들이 다양한 패지키 안에 존재합니다. 그러나, 이를 수행하는 기본 알고리즘은 동일하며, R 기본시스템만으로 작성이 가능합니다. 아래의 함수를 복사하여 사용하시면 됩니다.

```
fill <- function(x, first, last){
        n <- last-first+1
        for(i in c(1:length(first))) x[first[i]:last[i]] <- rep(x[first[i]], n[i])
        return(x)
}
```

2. 위에서 주어진 데이터에서 ID 변수에서 보이는 것처럼 같은 관측치가 여러번 반복 측정되어 ID가 반복적으로 입력이 되었을 때, SAS에서처럼 각 아이디별로 첫번째와 마지막 레코드를 알수 있는 .FIRST와 .LAST 같은 기능이 R에서는 어떻게 해야 하나요?

```
mydata$first <- !duplicated(mydata$ID)
mydata$last <- !duplicated(mydata$ID, fromLast=TRUE)
```

```
> mydata
   ID Week first  last
1   1   15  TRUE FALSE
2   1   NA FALSE FALSE
3   1   NA FALSE FALSE
4   1   NA FALSE  TRUE
```

```
5   2   18  TRUE FALSE
6   2   NA FALSE FALSE
7   2   NA FALSE FALSE
8   2   NA FALSE  TRUE
9   3   20  TRUE FALSE
10  3   NA FALSE  TRUE
```

3. 데이터의 일부분만 골라 내고 싶어요. 예를들면, 위에서 사용된 예제에서 ID 가 1과 2인 데이터만 골라내고 싶다면 아래와 같이 할 수 있습니다.

```
# 데이터 생성하기
mydata <- data.frame(ID=c(rep(1,4), rep(2,4), rep(3,2)), Week=c(15, NA, NA, NA, 18, NA, NA, NA, 20,

# ID 변수에 있는 ID를 기준으로 첫번째와 마지막 레코드의 위치 알아내기
idx.first <- which(!duplicated(mydata$ID))
idx.last <- which(!duplicated(mydata$ID, fromLast=TRUE))

# ID에 있는 NA값을 채워넣기
mydata$Week <- fill(x=mydata$Week, first=idx.first, last=idx.last)

# 개별 ID에 대한 첫번째와 마지막 레코드에 대한 논리값을 추가하여 데이터 확장하기
mydata$first <- !duplicated(mydata$ID)
mydata$last <- !duplicated(mydata$ID, fromLast=TRUE)

# 조건에 맞는 데이터 골라내기
select <- subset(x=mydata, subset=(ID %in% c(1,2)))
> select
  ID Week first  last
1  1   15  TRUE FALSE
2  1   15 FALSE FALSE
3  1   15 FALSE FALSE
4  1   15 FALSE  TRUE
5  2   18  TRUE FALSE
6  2   18 FALSE FALSE
7  2   18 FALSE FALSE
8  2   18 FALSE  TRUE

# 추가적인 조건 부여하기
select.1 <- subset(x=mydata, subset=( (ID %in% c(1,2)) & first==TRUE ))
> select.1
  ID Week first  last
1  1   15  TRUE FALSE
5  2   18  TRUE FALSE
```

4. wide format 데이터를 long format 으로 바꿀 수 있나요?

5. 여러개의 엑셀시트로 구성되어 있는 엑셀파일을 불러와 하나의 데이터셋으로 합치기

6. 가끔 리스트형으로 받아진 데이터가 중첩된 구조를 가지고 있어서, 한 번에 이를 불러오기를 해야할 때는 어떻게 해야할지.

7. `do.call()` 함수를 사용하는 법에 대해서..

8. `which.max()`와 `which.min()`을 사용하는 방법

9. list()과 data.frame()과의 관계
10. apply(), lapply(), sapply(), mapply()의 사용방법
11. 현재 패키지를

1.4 수치해석 및 시뮬레이션에 관련하여

1. 미분하기 – 어떤 예제가 좋을까? Binomial distribution 으로 제공해주기
2. 적분하기 – 몇 가지 예제가 있으면 좋을꺼 같음.
3. R에도 C와 같은 switch문이 존재하며, 그렇다면 어떻게 사용할 수 있나요?
4. warning(경고)와 error(에러)를 이용하는 법
5. try() 함수를 이용하여 에러를 컨트롤 해보기
6. tryCatch() 함수를 이용해서 에러를 컨트롤하기

```
result <- tryCatch(
{
        수행하고자 하는 표현식
},
warning = function(w) {
        위에서 수행한 표현식이 경고를 발생시킬때 어떻게 처리하고자 하는지에 대한 표현식
},
error = function(e) {
        위에서 수행한 표현식이 에러를 발생시킬때 어떻게 처리하고자 하는지에 대한 표현식
}, finally {
        위에서 수행한 표현식에 대한 최종적 처리를 위한 표현식
}
```

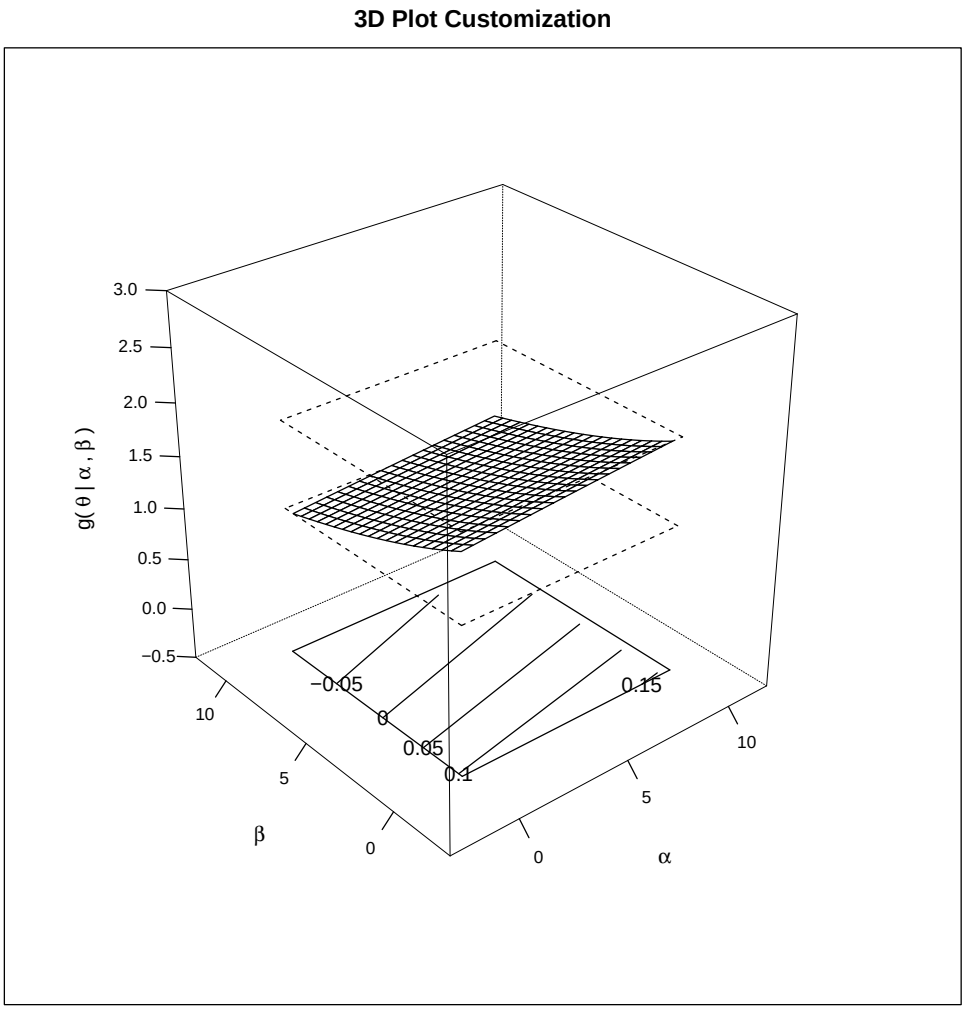
   예제는 내일 시간날때 작성

7. combn() 함수를 이용하여 모든 조합을 찾기
8. Metropolis-Hastings 알고리즘을 구현하는 프레임 워크 - 이것은 그냥 사용가능하게 바로 소스코드 붙여주기 (베이지안 컴퓨테이션에 많이 쓰임)
9. Newton-Raphson 알고리즘을 구현하는 방법 – optimization 에 관련된 일종의 설명도 추가해주면 좋을 것 같음
10. Laplace Approximation 알고리즘을 구현하는 방법 – 적분하는 방법에 많이 쓰임 (특히, 베이지안 컴퓨테이션)
11. Bootstrap 방법 – 요건 아주 좋은 패키지가 있음

1.5 비쥬얼라이제이션

1. coordinating system을 활용하기

2. Lattice 패키지를 이용하여 아래와 같은 그림을 생성해보기 (가장 단순한 예제임 - 팁 보다는 튜토리얼 형식으로?)



Lattice Example 1

3. LaTeX의 문서에 포함될 .eps 그래픽을 R에서 뽑았을 때는 아무런 문제가 없어 보였는데, 정작 pdf로 문서를 뽑아 보니까 이 그래픽이 들어간 페이지가 90도로 돌아가 있거나 혹은 그래픽이 90도로 회전되어 있을 경우에는 아래와 같이 하면 됩니다.

```
postscript(file=``filename.eps'', onefile=FALSE, horizontal=FALSE)
```

이 문제에 대한 출처는 postscript 도움말입니다.

이문제를 다른 방법으로도 해결할 수 있습니다. (대충 서너개 더 있음).

4. 새로운 그래픽 객체를 생성하는 방법을 설명해줘서 사용자가 추후에 독립적인 그래픽을 생성할 수 있게 도와주기

1.6 데이터 입출력 및 파일관리 유틸리티

1. `read.table` 계열의 함수를 이용하여 데이터를 불러올 때 첫번째 인자는 파일의 위치와 파일명이 입력된 문자열이어야 합니다. 그런데, 간혹 문법에서 틀린 점도 없고, 불러오고자 하는 데이터 파일도 올바른 파일경로에 위치하고 있음에도 불구하고, 데이터를 찾을 수 없다고 하는 경우가 있습니다. 이것은 내부적으로 파일경로에 띄어쓰기, 특수문자, 혹은 특수한 인코딩 등 다양한 이유로 인하여 파일경로가 올바르게 처리되지 않았기 때문입니다. 아래와 같은 방법으로 `read.table()` 함수 사용시 `file.choose()` 함수를 함께 사용하면 이러한 문제를 해결이 가능합니다.

```
mydata <- read.table(file.choose(), header=TRUE, sep=",")
```

1.7 클래스와 메소드 그리고 패키지 제작

1. 패키지를 만들고 싶어요. (흠.. Generalized Linear Model 프레임워크 흉내내서 똑같이 만들어보기 실습자료로 제공해주기)

1.8 인코딩과 한글

1. 불러오고자 하는 데이터의 인코딩이 UTF-8가 아닐때 이를 확인하고, 데이터를 올바르게 불러오기 위한 내용은 `http://lists.r-forge.r-project.org/pipermail/ihelp-urquestion/2013-April/000017.html` 를 읽어보시길 바랍니다.

2. `R`을 한국어가 아닌 영문으로 사용하고 싶습니다 (버전에 관계없이 일반적으로 통용되는 방법 - 윈도우즈 사용자에게 맞추어 작성됨). 이를 설정하는 방법에 대해서는 `http://lists.r-forge.r-project.org/pipermail/i` 을 읽어보시길 바랍니다.

1.9 간단한 GUI 제작 해보기

1. 다른 언어로 인터페이싱 하는 방법마로, 그냥 `R`에서 주어지는 패키지를 이용해서 간단한 GUI 환경만들기

2. 아마도... R Commander를 확장하는 방법을 예로 들면 좋을 것 같음

3. 원리도 간단히 설명해주면 더욱 좋을 것 같음.

1.10 답변되지 않을 수도 있는 질문들

1. (접수: 2013-04-18, Reproducibility=NA) R의 장점이자 단점이라고 생각되는 것 중에 하나가 엄청난 수의 패키지들임. 즉 어떤 분석을 하고자 할 때 그것에 대해 하나의 패키지가 있는 것이 아니라 대체적으로 사용가능한 패키지들이 존재하는데 이들 중 어느 것을 써야할 지 잘 모름. 다른 분석 프로그램의 경우 이러한 문제가 없는데... 결국엔 어떻게 제일 성능이 좋은? 결과가 신뢰할 만한? 좋은 패키지를 선택하는가를 알려주었으면 좋겠씀돠.

   (답변) 이것은 경험에 해당되며, 해당분야의 전문가로부터의 조언을 받는 것이 안전합니다. 그렇지 않다면, 직접 베이스를 이용하여 작성하면 됩니다.