한국한의학연구원, 구본초

통계 프로그래밍 언어

Contents

List of Tables

List of Figures

Course Overview

본 문서는 2020년도 1학기 정보통계학과에서 개설한 "통계 프로그래밍 언어" 강의를 위해 개발한 강의 노트이며, Yihui Xie가 개발한 **bookdown** 패키지 (Xie, 2019)를 활용하여 생성한 문서이고 Google Chrome 또는 Firefox 브라우저에 최적화 됨. 아울러 충남대학교 정보통계학과 이상인 교수님의 2019년도 2학기 "통계패키지활용" 강의 노트와 동국대학교 ICT빅데이터학부 김진석 교수님의 R 프로그래밍 및 실습[1] 강의 자료 내용을 참고함.

본 강의 노트는 주 단위로 업데이트될 예정임.

강의소개

R은 뉴질랜드 오클랜드 대학의 Robert Gentleman 과 Ross Ihaka 가 AT&T 벨 연구소에서 개발한 S 언어를 기반으로 개발한 GNU 환경의 통계 계산 및 프로그래밍 언어이다. 현재 R 소프트웨어는 통계학 뿐 아니라 데이터 과학을 포함한 의학, 생물학 등 다양한 분야에서 활용되고 있으며 특히 통계 소프트웨어 개발과 데이터 분석에 많이 활용되고 있다. 본 강의는 데이터 분석을 위한 R의 기초 문법과 통계학 입문에서 학습한 몇 가지 중요한 통계적 이론에 대한 시뮬레이션 방법을 다룬다. 아울러 R package를 활용한 데이터 핸들링 및 시각화 그리고 Rmarkdown을 활용한 재현가능(reproducible)한 문서 작성법에 대해 학습하고자 한다.

교과 목표

- **R 기초 문법 습득**
- **R package를 활용한 데이터 핸들링 및 자료 시각화**
- **R 시뮬레이션을 통한 통계학 기초 이론 확인**
- **R을 이용한 데이터 분석 실습**
- **R markdown을 이용한 재현가능(reproducible)한 보고서 작성 방법 습득**

선수과목

통계학 개론

수업 방법

- **강의: 50 %**
- **실험/실습: 50%**

평가방법

- **중간고사: 40 %**
- **기말고사: 40 %**
- **출석: 10 %**
- **과제: 10 %**

수업 규정

- 3번 지각은 1번 결석으로 처리
- 특별한 사유 없이 수업 중간에 이탈한 경우 결석으로 처리
- 특별한 사유로 인해 결석 또는 지각을 할 경우 사유를 증빙할 수 있는 서류 제출 시 출석으로 인정
- 출결 미달, 중간 또는 기말고사 미 응시인 경우 F 학점으로 처리
- 수업 중 휴대폰 및 각종 모바일 기기 사용 금지

교재 및 참고문헌

별도의 교재 없이 본 강의 노트로 수업을 진행할 예정이며, 수업의 이해도 향상을 위해 아래 소개할 도서 및 웹 문서 등을 참고할 것을 권장함.

참고 자료

- 실리콘밸리 데이터과학자가 알려주는 따라하며 배우는 데이터 과학 (권재명, 2017)
- R을 이용한 데이터 처리&분석 (서민구, 2014)
- R 그래픽스 (유충현 et al., 2005)
- ggplot2: elegant graphics for data analysis[2] (Wickham, 2016)
- R for data science[3] (Wickham and Grolemund, 2016)
- Statistical Computing with R (Rizzo, 2019)

[2] https://ggplot2-book.org/
[3] https://r4ds.had.co.nz/

강의 계획

TABLE 0.1: 강의 계획표

주차	강의 내용	과제
Week 1	R 소개, R/R Studio 설치, R 패키지 설치, R 맛보기 및 markdown 문서 만들기	과제 1
Week 2	R 자료형 : 스칼라, 벡터, 리스트	
Week 3	R 자료형 : 행렬 및 배열	과제 2
Week 4	R 자료형 : 팩터, 테이블, 데이터 프레임	
Week 5	R 자료형 : 문자열과 정규 표현식	과제 3
Week 6	데이터 프레임 가공 및 시각화 I	
Week 7	데이터 프레임 가공 및 시각화 II	과제 4
Week 8	중간고사	
Week 9	데이터 프레임 가공 및 시각화 III	
Week 10	R 프로그래밍 : 조건문, 반복문, 함수	과제 5
Week 11	통계시뮬레이션 I: 표본분포 및 중심극한정리	
Week 12	통계시뮬레이션 2: 신뢰구간과 가설검정	과제 6
Week 13	R을 이용한 기초통계 분석	
Week 14	R markdown 활용	과제 7
Week 15	기말고사	

Part I

Get Started

1

Introduction

1. R프로그램

- 데이터 분석을 위한 자료 전처리, 통계 및 시각화를 지원하는 컴퓨터 언어 및 환경
- 1980년 AT&T 벨 연구소의 John Chambers가 개발한 S 언어를 기반으로 1995년 뉴질랜드 Auckland 대학의 통계학과 교수 Robert Gentleman과 Ross Ihaka 가 개발
- GNU[1] 기반의 오픈 소스
- 통계학, 전산학, 생물학, 의학 등 거의 모든 학문분야에서 분석도구로 활용되고 있고, 최근 data science 분야에서 널리 활용

2. R 언어의 특징

- **무료 소프트웨어**
- CRAN (Comprehensive R Archive Network)[2]에서 배포
- 특정 vendor가 아닌 전 세계 연구자들이 개발한 알고리즘 및 최신 함수 활용 가능(packaging system)
- 범용적으로 사용되는 거의 대부분의 운영체제(Windows, Mac, Linux)에서 작동 가능

[1] https://en.wikipedia.org/wiki/GNU_Project
[2] http://cran.r-project.org/web/view

- 방대한 개발 및 사용 생태계 형성
- 강력한 그래픽 기능

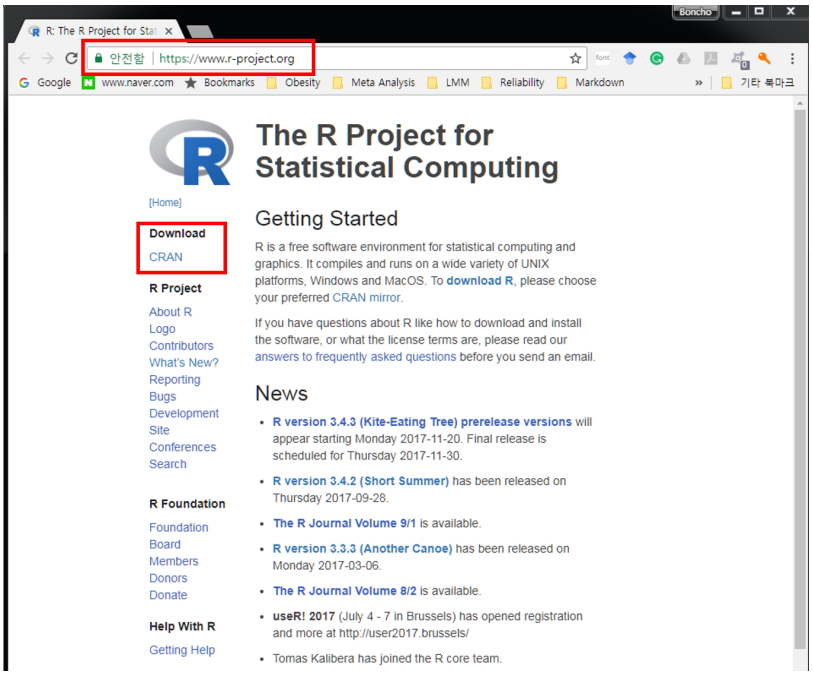
유용한 웹 사이트: R과 관련한 거의 모든 문제는 Googling (구글을 이용한 검색)을 통해 해결 가능(검색주제 + "in R" or "in R software")하고 많은 해답들이 아래 열거한 웹 페이지에 게시되어 있음.

- R 프로그래밍에 대한 Q&A: Stack Overflow[3]
- R 관련 웹 문서 모음: Rpubs[4]
- R package에 대한 raw source code 제공: Github[5]
- R을 이용한 통계 분석: Statistical tools for high-throughput data analysis (STHDA)[6]

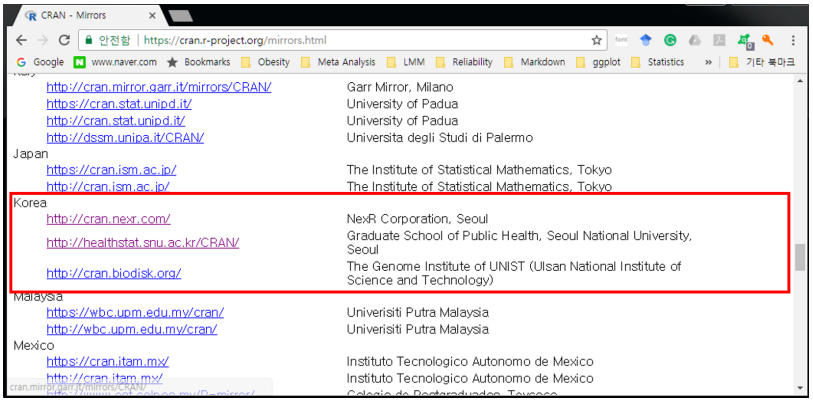
1.1 R 설치하기

R 다운로드 사이트: `https://www.r-project.org` 또는 `https://cran.r-project.org`

1. 웹 브라우저(i.e. Explore, Chrome, Firefox 등)의 주소 입력창에 `https://www.r-project.org`
2. 좌측 R Logo 하단 Download 아래 CRAN 클릭

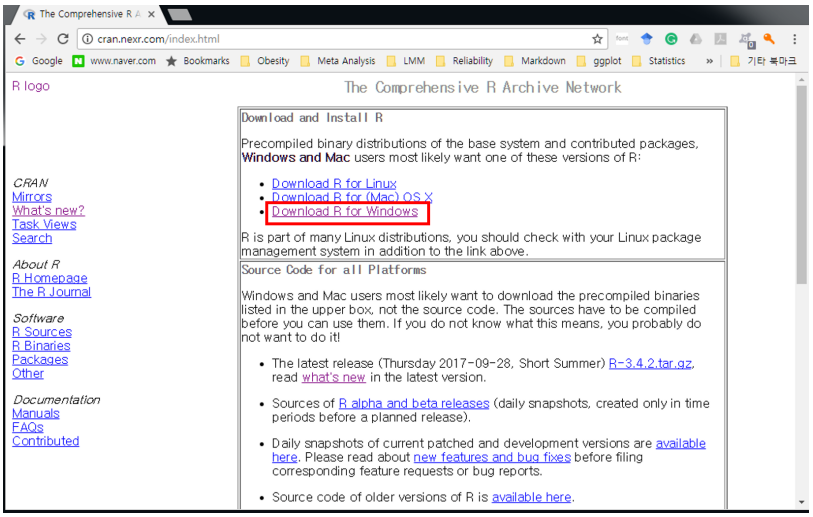


3. 클릭 후 연결한 페이지를 스크롤 후 Korea 아래 링크[7] 클릭

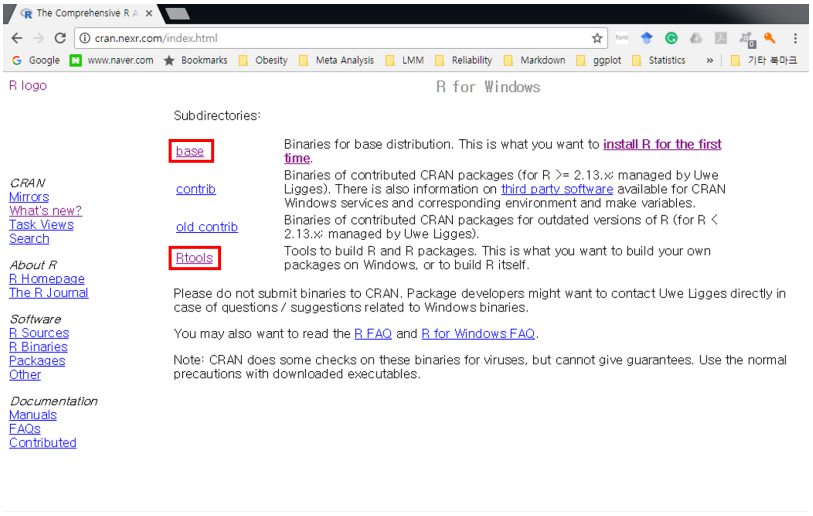


4. 클릭 후 세 가지 운영체제(Linux, Mac OS X, Windowns)에 따른 R 버전 선택 가능[8]

[7] 해당 링크들은 접속 시점에 따라 변경될 수 있음
[8] 본 노트는 Windows 버전 설치만 다룸



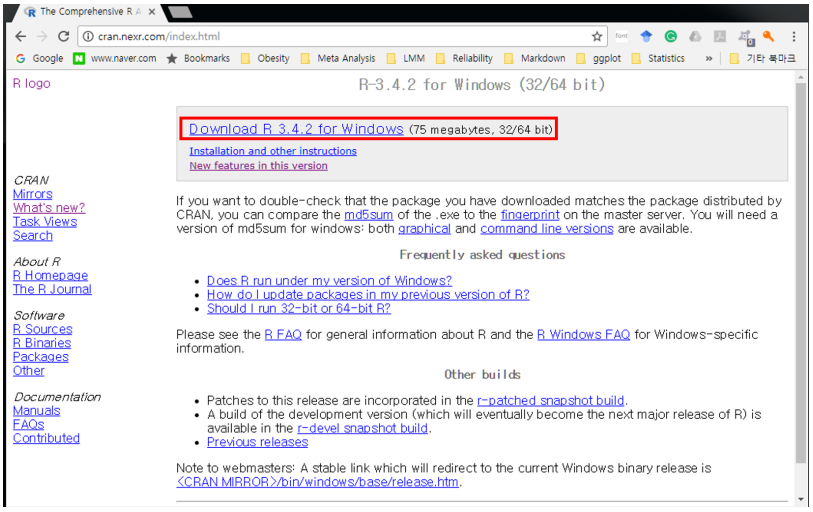
5. **Downloads R for Windows** 링크 클릭하면 다음과 같은 화면으로 이동



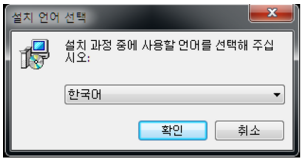
다음 하위폴더에 대한 간략 설명

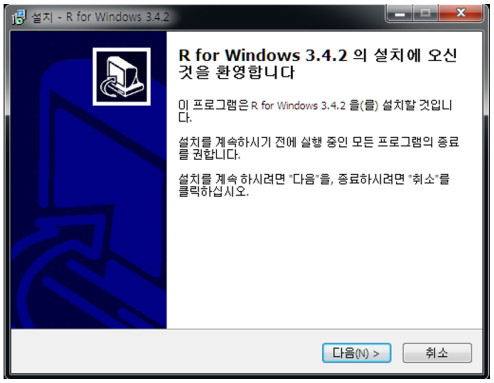
- **base**: R 실행 프로그램
- **contrib**: R package의 바이너리 파일
- **Rtools**: R package 개발 및 배포를 위한 프로그램

6. 위 화면에서 **base** 링크 클릭 후 아래 화면에서 **Downloads R 3.x.x for Windows** 를 클릭 후 설치 파일을 임의의 디렉토리에 저장 및 실행

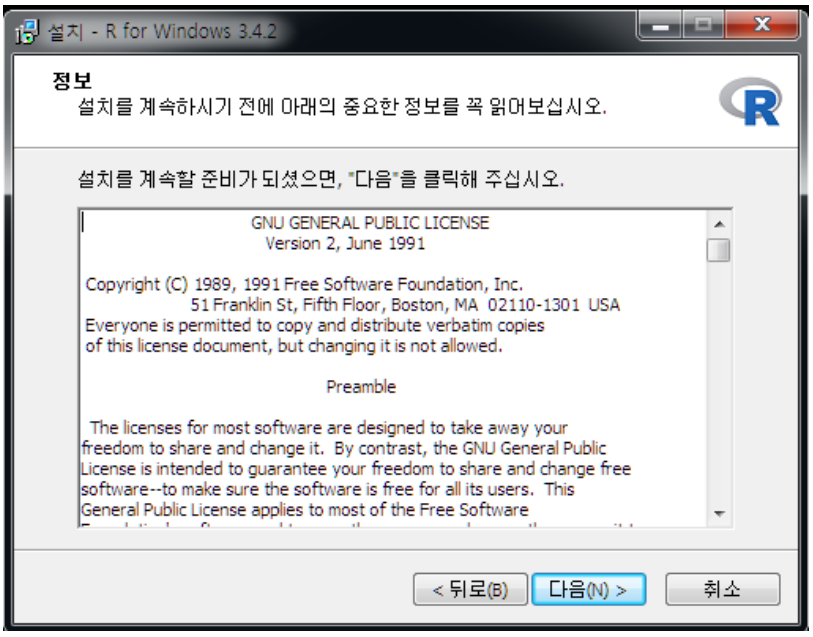


7. 다운로드한 파일을 실행하면 아래와 같은 대화창이 나타남
   - 한국어 선택 → 환영 화면에서 [다음(N)>] 클릭

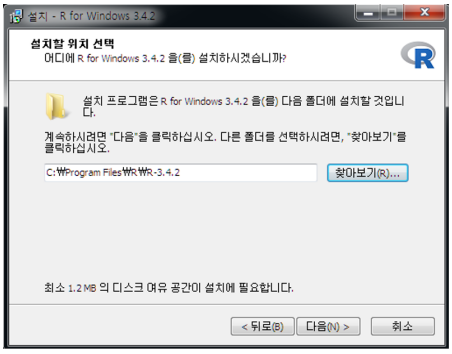


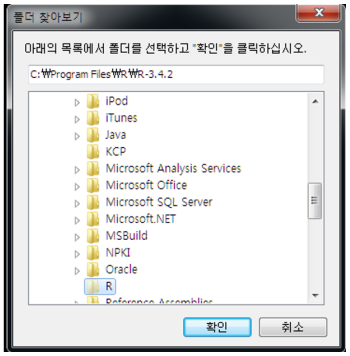


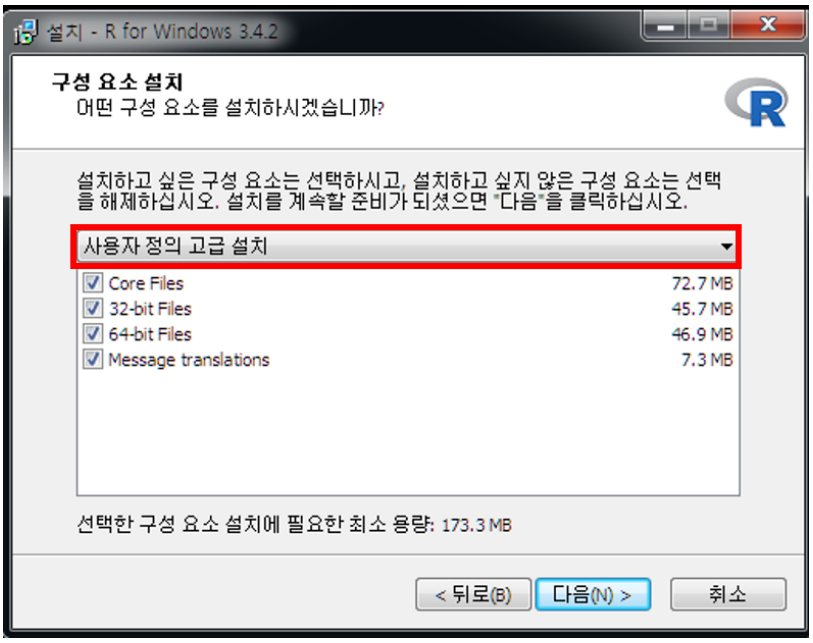
8. GNU 라이센스에 대한 설명 및 동의 여부([다음(N)>]) 클릭



9. 설치 디렉토리 설정 및 구성요소 설치 여부
   - 원하는 디렉토리 설정 (예: `C:\R\R-3.x.x`)
   - 기본 프로그램 ("Core Files"), 32 또는 64 bit 용 설치 파일, R console 한글 번역 모두 체크 뒤 [다음(N)>] 클릭

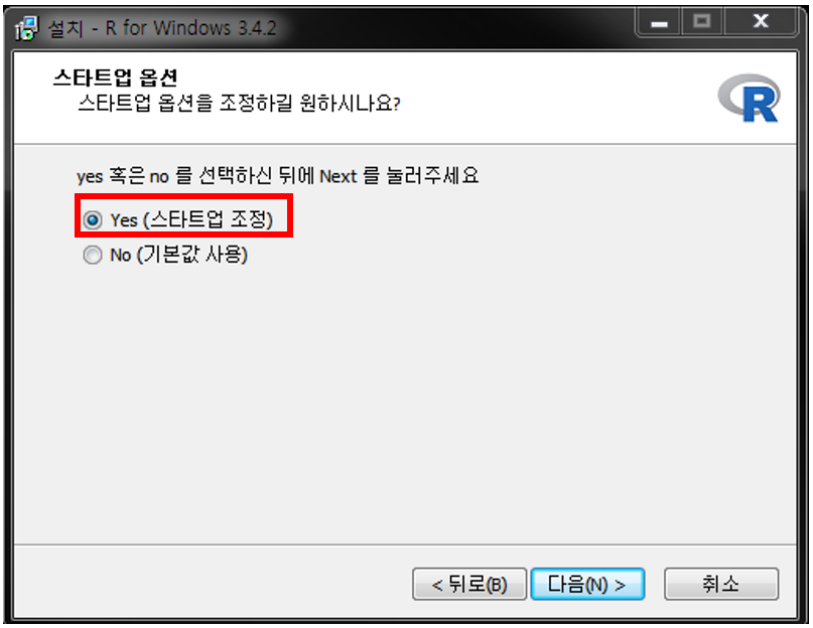






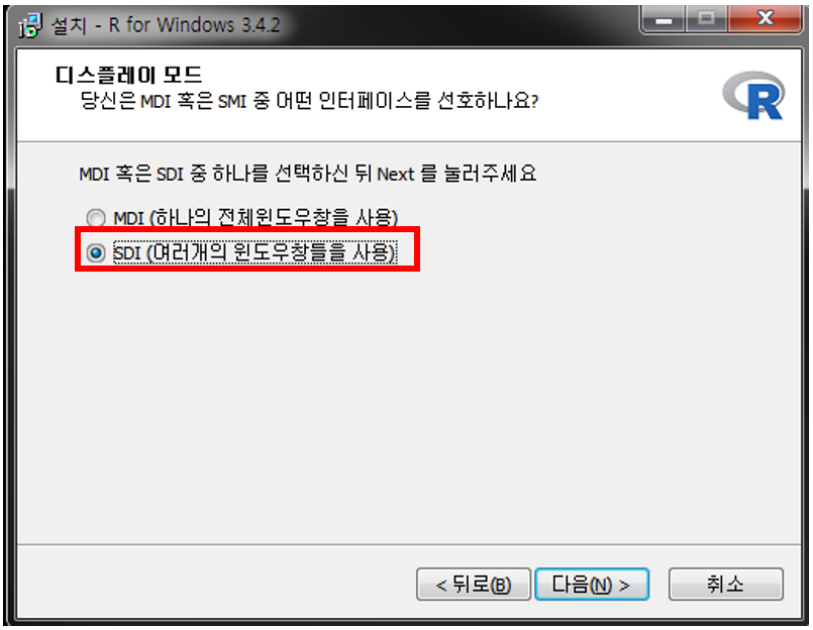
10. R 스타트업 옵션 지정

- 기본값(“No” check-button)으로도 설치 진행 가능
- 본 문서에서는 스타트업 옵션 변경으로 진행

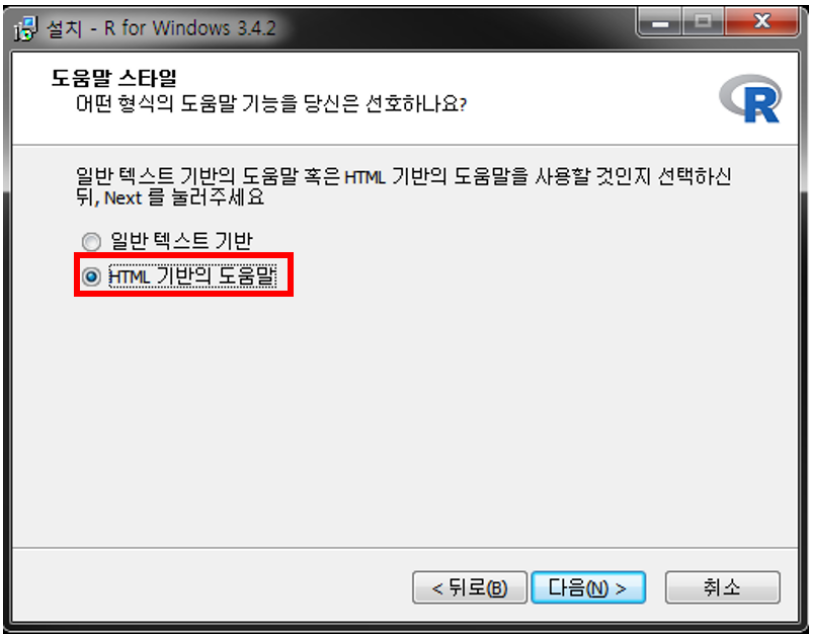


11. 화면표시방식(디스플레이 모드) 설정 변경

- MDI: 한 윈도우 내에서 script 편집창, 출력, 도움말 창 사용
- SDI: 다중 창에서 각각 script 편집창, 출력, 도움말 등을 독립적으로 열기

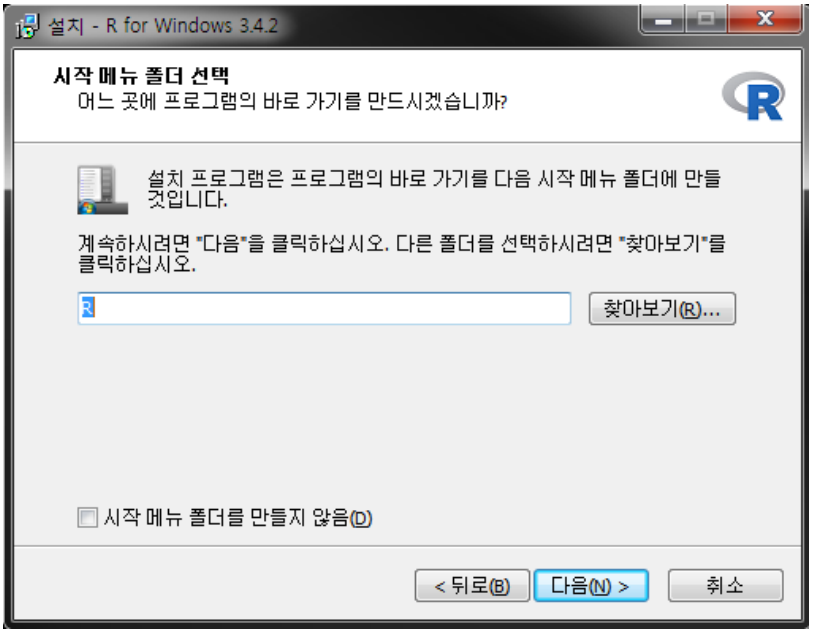


12. 도움말 형식에서 HTML 도움말 기반 선택



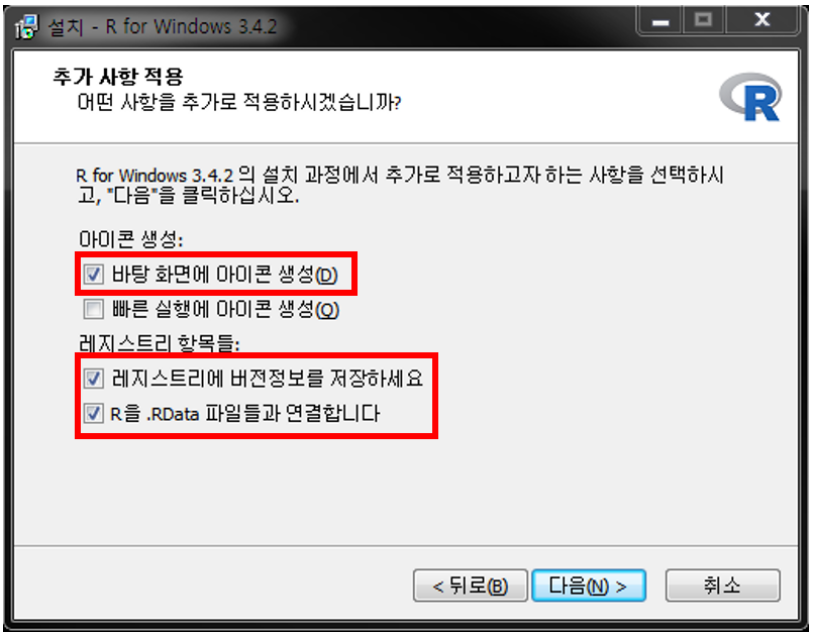
13. 시작메뉴 폴더 선택

- “바로가기”를 생성할 시작 메뉴 폴더 지정 후 [다음(N)>] 클릭 후 설치 진행
- 하단 “시작메뉴 폴더 만들지 않음” 체크박스 표시 시 시작메뉴에 “바로가기” 아이콘이 생성되지 않음(실행에 전혀 지장 없음)



14. 추가 옵션 지정 : 바탕화면 아이콘 생성 등 추가적 작업 옵션 체크 후 [다음(N)>] 클릭 → 설치 진행

- 설치된 R 버전 정보 레지스트리 저장 여부
- `.Rdata` 확장자를 R 실행파일과 자동 연계



15. 설치 완료 후 바탕화면의 R 아이콘을 더블클릭하면 Rgui가 실행

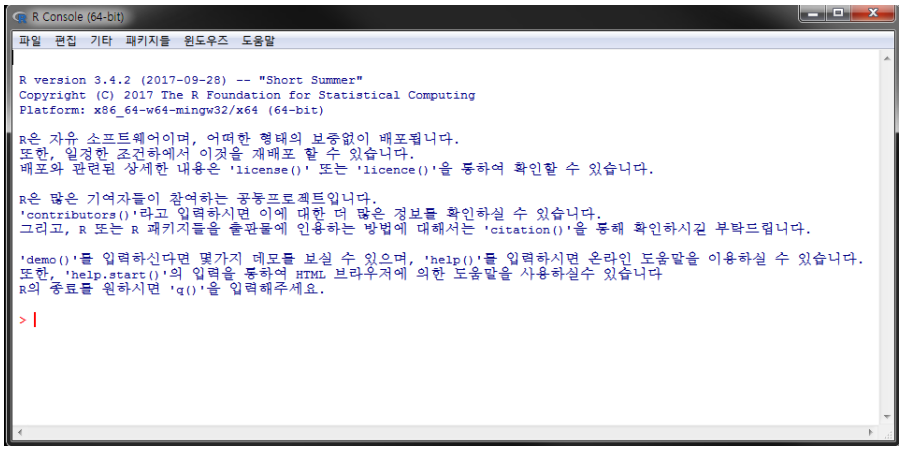


FIGURE 1.1: Windows에서 R 실행화면(콘솔 창, SDI 모드)

1.2 R 시작 및 작동 체크

실습: 설치된 R을 실행 후 보이는 R 콘솔(consle) 창에서 명령어를 실행하고 결과 확인

그림 Figure 1.1 에서 > 기호는 R의 명령 프롬프트(prompt) 임

1. 문자열 출력

```
#문자열 출력
print("Hello R") #문자열
```

```
[1] "Hello R"
```

기호는 주석의 시작을 의미하고 실제로 실행되지 않음 같은 행에서 # 뒤 내용의 코드 역시 실행되지 않음

2. a 라는 변수에 숫자 9, b라는 변수에 숫자 7를 할당 후 출력

```
# 수치형 값(scalar)을 변수에 할당(assign)
# 여러 명령어를 한줄에 입력할 때에는 세미콜론(;)으로 구분
a = 9; b = 7
a
```

```
[1] 9
```

```
b
```

```
[1] 7
```

3. 변수 a와 b의 사칙연산

```
a+b; a-b; a*b; a/b
```

```
[1] 16
```

```
[1] 2
```

```
[1] 63
```

```
[1] 1.285714
```

4. R 그래픽 맛보기 : 정규분포로부터 난수 100개 생성 후 생성된 데이터에 대한 히스토그램 작성

```
# 난수 생성 시 값은 매번 달라지기 때문에 seed를 주어 일정값이 생성되도록 고정
# "="과 "<-"는 모두 동일한 기능을 가진 할당 연산자임
#평균이 0이고 분산이 1인 정규분포에서 난수 100개 생성
set.seed(12345) # random seed 지정
x <- rnorm(100) # 난수 생성
hist(x) # 히스토그램
```

R 명령어 또는 전체 프로그램 소스 실행 시 매우 빈번히 오류가 나타나는데, 이를 해결할 수 있는 가장 좋은 방법은 앞에서 언급한 Google을 이용한 검색

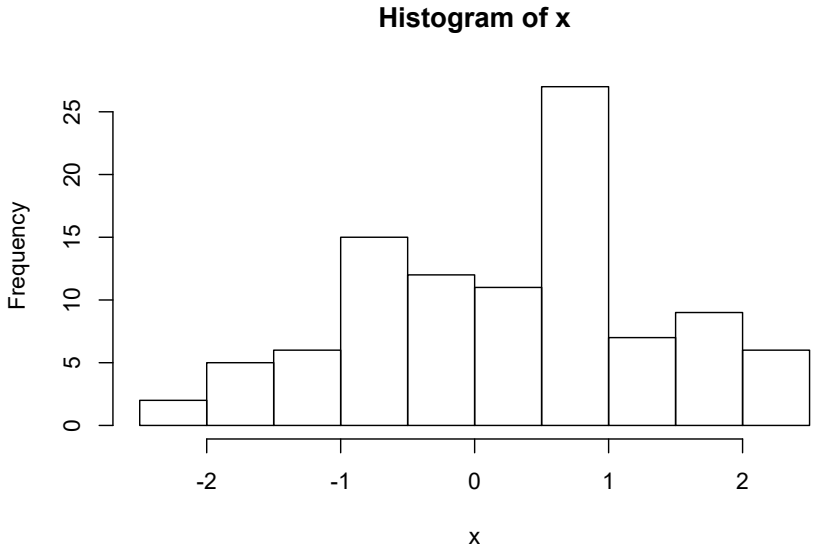


FIGURE 1.2: 정규분포 100개의 히스토그램

또는 R 설치 시 자체적으로 내장되어 있는 도움말을 참고하는 것이 가장 효율적임.

```
help
```

Bibliography

Rizzo, M. L. (2019). *Statistical computing with R*. CRC Press.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Xie, Y. (2019). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.16.

권재명 (2017). 실리콘밸리 데이터 과학자가 알려주는 따라하며 배우는 데이터 과학. 제이펍, 1st edition. ISBN 979-1185890869.

서민구 (2014). *R*을 이용한 데이터 처리&분석. 길벗, 1st edition. ISBN 978-8966188260.

유충현, 이상호, and 김정일 (2005). *R* 그래픽스. 자유아카데미, 1st edition. ISBN 978-8973385539.