Chapter 1

완전 초보에요

이 문서에서 "초보"라는 의미는 아래에 나열된 사항들중 두 가지 이상에 해당되시는 분들을 의미합니다.

- R 이라는 프로그래밍 언어 이전에 다른 프로그래밍 언어에 대한 경험이 전무하신 분,
- 유닉스와 리눅스 시스템에 익숙하지 않으신 분,
- 기초 통계 분석에 대한 도움이 필요하신 분,
- 그냥 무엇을 해야할지 막막함에 쌓여 계신분

이에 해당하시는 분들은 꼭 "사용전 반드시 알아야 할 7가지 숙지사항" 섹션을 꼭 읽어주시길 부탁드립니다. R을 사용하시는데 있어 숙지사항의 내용을 기억하신다면, 매우 도움이 될 것입니다.

1.1 꼭 먼저 알아야 할 7가지

초보라고 생각하시는 분들께서는 아래의 내용들에 미리 알고 계시면 R을 사용하는데 도움이 됩니다.

데이터 입력과 처리: R 에서는 이용되는 모든 데이터들에 대한 처리는 열방향으로 이루어집니다. 이 말의 뜻은 데이터의 입력 및 변형, 그리고 연산에 사용되는 데이터들에 대한 처리순서는 열방향으로 나열된 후에 이루어지는 것을 말합니다. 예를들어, 1부터 12까지 12개의 정수로 이루어진 수열 (즉, $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$)을 R은 아래와 같이 이해합니다. 열방향으로 이루어진다고 하더라도 수학적인 표현은 동일합니다(즉, 4행 3열 또는 4x3 행렬).

```
1  5   9
2  6  10
3  7  11
4  8  12
```

따라서, 이 행렬에서 5라는 숫자값은 행렬의 1행 2열에 위치하고 있다고 할 수 있으며, 행렬의 5번째의 값이라고도 할 수 있습니다. 5번째라고 하는 것은 R이 열기준으로 데이터를 인식하는 이유 때문입니다(만약 어떤 프로그램이 행으로 데이터를 인식한다면 5번째의 값은 6이 됩니다).

R의 환경설정 및 유틸리티: R을 잘 사용하고 싶은 마음은 굴뚝같은데 생각보다 뜻대로 되지 않을 것입니다. 그 이유는 R에 익숙하지 않아서 입니다. 익숙하다는 말은 다양한 방향으로 쓰이나, R은 어떤 분석을 위해 그 분석이 가진 프로세스를 진행하고자 하는데 사용되는 도구입니다. 따라서, 도구를 얼마나 잘 쓰는가가 어떤 분석을 얼마나 효율적으로 할 수 있는가와도 같은 말이 됩니다. 대부분의 경우 초보자는 R의 환경설정에 관련된 부분이 어떻게 R 프로그래밍에 영향을 미치는지 모릅니다. 따라서, R에 빨리 익숙해지고 싶으신 분들은 **환경설정과 유틸리티**라는 챕터를 먼저 훑어보시길 바랍니다.

대화식 사용을 통한 분석결과를 확인하는 방법에 대해서: R은 사용자가 주어지는 업무를 시키는 대로만 수행하는 유용한 프로그램일뿐, 그 이상의 내용은 수행하지 않는다는 점을 반드시 명심하셔야 합니다. 따라서, R 프로그램을 시작하게 되면 아래와 같은 기호로 표시되는 프롬프트 (즉, 사용자의 명령어를 기다리는 기호)를 보여주게 됩니다. 모든 명령어는 > 기호 뒤에 작성하게 됩니다. 가장 중요한 것은 분석자가 머릿속으로 상상하거나 기대하는 업무가 있다면, 분석자는 반드시 그 업무가 이루어지는 프로세스에 대해서 잘 알고 있어야 합니다. 따라서, R은 다른 통계 소프트웨어들과 같이 분석된 결과를 미리 보여주거나 혹은 분석이 된 모든 결과를 한 번에 다 보여주지 않습니다. 분석자가 확인하고자 하는 결과를 R에서 제공하는 함수를 통하여 중간 결과물들을 확인할 수 있습니다. 우리는 분석자가 작업을 어떻게 진행하고 확인할 것인가에 대한 프로세스 차트를 먼저 작성한뒤 분석을 수행하기를 권장합니다.

통계분석의 절차: 분석이라 것은 데이터에 대한 이해를 통하여 데이터가 가진 특징을 수학적 표현으로서 설명하기 위한 과정입니다. 따라서, 데이터에 대한 직관적인 이해를 위해서 시각화 작업과 통계 모형이 요구하는 데이터 형식을 만드는 것이 중요합니다. 이를 전처리 과정이라고 하며, 통계분석은 크게 아래와 같은 절차를 밟아 이루어집니다.

1. 데이터 입출력과 클리닝, 그리고 분석 전처리 관련 테크닉들
2. 분석전 탐색적 시각화 작업
3. 통계모형의 결정 및 적용
4. 모형 적용 후의 보고서 생성 및 시각화 작업
5. 통계 모형 자체의 개발 또는 자동화 시스템 구축.

따라서, 이 문서는 위에서 설명한 과정에 해당하는 순서대로 챕터들을 구성하였습니다.

한글 표현과 인코딩에 관련하여: R은 한국어 사용자를 위한 한국어 인터페이스를 지원하고 있습니다. 그러나, 사용자는 이러한 한국어 지원이 단순히 사용자의 편의를 위한 선택적 사항이라는 점을 반드시 알고 있어야 합니다. 또한, 분석시 본래의 정교한 표현은 번역된 한국어 보다는 원래의 영문이라는 점도 잊지 마셔야 합니다. 현재 한국어에 관련된 작업은 UTF-8이라는 인코딩에 기반하여 한국어 품질관리 프로그램 (`http://ihelp.r-forge.r-project.org/lang_msg.html`)을 통하여 이루어 지고 있으나, R을 한국어가 아닌 영문으로 설정하기 위해서는 다음의 주소를 눌러 그 내용을 확인해주시길 부탁드립니다. 이 내용은 버전에 관계없이 일반적으로 통용되는 방법이나 윈도우즈 사용자에게 맞추어 작성되었습니다.

`http://lists.r-forge.r-project.org/pipermail/ihelp-urquestion/2013-April/000003.html`

본래 데이터를 자국의 언어로 표현하는 방법은 UTF-8 이라는 인코딩을 이용하므로, 간혹 데이터가 올바르게 보이지 않을 경우가 있습니다. 이러한 문제를 해결하기 위해서는 "R을 지원하는 인코딩 중 올바르게 한국어를 표현해주는 코드를 찾는 방법" (`http://lists.r-forge.r-project.org/pipermail/ihelp-urquestion/2013-April/000017.html`) 를 읽어보시길 바랍니다.

간혹, 잘못된 한글처리가 시스템 에러를 야기하는 경우가 있으므로, 이러한 경우에 한국어로 R을 사용하시는 분들께서는 다소 번거로우실지라도 보다 안정적이고 나은 R을 제공하기 위하여 그 내용을 ihelp-urquestion@lists.r-forge.r-project.org 주소로 이메일을 보내주신다면 감사하겠습니다.

단, 이러한 한글처리는 분석에 연관된 수치연산과는 아무런 관계가 없음을 반드시 알아주시길 부탁드립니다.

대소문자 구분에 관련하여: 많은 분들이 이전에 SAS 소프트웨어를 사용하여 분석을 수행하셨을 것입니다. SAS에서는 대소문자를 구분하지 않고 프로그램을 작성하게 되지만, R에서는 대소문자를 구분하므로 A 라는 변수와 a 라는 변수는 서로 다른 것임을 명심하시길 부탁드립니다.

객체: R에서 다루어지는 모든 것을 객체라고 합니다.

함수의 사용에 관련하여: R보다도 일반적으로 프로그래밍을 다소 익숙하게 다룬다는 것은 하고자 하는 업무에서 어떤 함수가 적재적소에 쓰여야 할 지를 아는 것입니다. 따라서, 제공되어는 함수가 무엇이 있으며 어떤 함수가 언제 어떻게 사용되는가를 알고 있는 것은 작업하는데 도움을 줍니다. 함수를 사용할 때, R은 지시된 인자(named argument)와 함께 사용하는 것이 좋습니다.

패키지와 관련하여: R은 Add-on이라는 패키지 시스템을 이용합니다. 이것은 기본 베이스 시스템에 추가로 필요한 기능들을 추가하는 의미입니다. 이러한 내용을 잘 모르는 상태에서 초보자가 가장 많이 겪는 실수는 어떤 함수를 사용하고자 할 때 "xxx 함수가 없습니다" 또는 "xxx 함수를 찾을 수 없습니다"입니다. 이는 사용하고자 하는 함수가 R 기본 배포판에 포함되어 있지 않은 어떤 사용자에 의해서 제공된 특정한 패키지내에서 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)
```

패키지의 설치, 확인에 관련된 사항은 "통계모형의 선택과 적용"이라는 챕터에 기록해두었습니다.

1.2 왜 R을 사용하나요?

R을 사용하는 이유는 아마도 아래와 같은 이유이기 때문입니다. (R Documentation에는 없는 이 문서의 지은이 개인의 생각임을 명심하시길 바랍니다)

1. R은 매우 다양한 분야에서 개발되고 적용되는 최신 통계기법을 적용할 수 있는 자유소프트웨어이기 때문입니다.

2. 행렬기반의 객체지향적 프로그래밍 언어이기 때문입니다.

3. 다른 소프트웨어들에 비교하여 문법적 사용의 자유롭기 때문일 것입니다.

1.3 통계소프트웨어의 종류

R이라는 통계소프트웨어를 대체할 수 있는 다른 소프트웨어들은 다음과 같습니다.

- S-PLUS (상업용 버전의 S 언어 소프트웨어)
- MATLAB (R과 같은 행렬기반의 언어)
- SPSS
- Octave (MATLAB의 GNU 버전)
- Python (프로그래밍 언어)
- SAS
- STATA