

R Commander 사용하기 *

John Fox

Version 1.9-1 (4 September 2012)

1 R Commander 시작하기

이 문서는 윈도우즈(Windows)용 R에서 R Commander의 사용 설명을 중심으로 기술한다. 애플 맥(Mac) OS X와 리눅스(Linux), 유닉스(Unix) 시스템에서 R Commander의 모습과 사용법에는 작은 차이가 있다. 이 플랫폼들에서 R Commander를 설치하는데 필요한 정보는 R Commander의 웹페이지 <<http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/index.html>>, 또는 <tinyurl.com/Rcmdr>의 설치 설명서에서 직접 이용가능하다.

R이 작동할 때, R Console에서 `library(Rcmdr)`라고 명령어를 입력하는 것으로 Rcmdr 패키지를 불러오면 R Commander의 그래픽 사용환경("GUI")이 시작된다. 윈도우즈에서 최상으로 작동하려면, R Commander는 단일-문서 인터페이스 (SDI)로 사용되어야 한다.¹ Rcmdr 패키지를 불러온 이후, R Console과 R Commander 윈도우즈는 그림1과 그림2처럼 보여야 한다. 이 스크린의 이미지는 윈도우즈 7에서 만들어졌다; 만약 다른 윈도우즈 버전 (또는, 물론, 다른 컴퓨팅 플랫폼)을 사용한다면, 스크린의 모양새가 다를 수 있다.²

R Console과 R Commander 윈도우즈는 데스크탑에서 자유롭게 움직일 수 있다. 데이터셋을 읽고, 조작하고 분석하기 위해서 R Commander의 대화상자와 메뉴를 이용할 것이다.

- R Commander GUI에 의해서 만들어진 R 명령어는 R Commander 안에서 (스크립트 창 이라고 이름붙여진) 위쪽의 텍스트 창 에 보인다. 당신은 스크립트 창이나 R Console의 > 프롬프트에 직접 R 명령어를 입력할 수도 있다; 하지만, R Commander의 주된 목적은 명령어를 손으로 입력하는 것을 피하는데 있다.
- 인쇄된 출력물은 기본설정으로 (출력창 이라고 이름붙여진) 두 번째 텍스트 창 에 보인다.
- 아래의 회색 창은 (Messages 라고 이름붙여진) 오류 메시지, 경고문, 일부 다른 정보 ("notes"), 예를 들어 그림 2에서의 시작 메시지를 보여준다.
- 당신이 그래프를 출력하면, 이것들은 독립된 Graphics Device 창에 보인다.

*fox (2005)에 기초해서 작성 및 업데이트되었습니다. jfox@mcmaster.ca로 연락하십시오. [역자주석: 이 글의 원자료는 Getting-Started-with-the-Rcmdr.pdf 입니다. 한글번역문에 대한 문의는 신중화(shin.jonghwa@r-project.kr)로 연락하십시오.]

¹윈도우즈 버전의 R은 일반적으로 R Console 창, 세션 중간에 생성되는 Graphical Device 창 및 R의 처리과정과 관련된 다른 창들을 포함하는 다중-문서 인터페이스 ("MDI")에서 작동한다. 반면에 단일-문서 창("SDI")에서 R Console, Graphical Device 창들은 마스터 창에 포함되지 않는다. R을 SDI 환경에서 작동시키는 여러 방법 있다 - 예를 들어, R이 설치될 때, R의 etc 하위디렉토리에 있는 Rconsole 파일을 편집하거나, 또는 R의 데스크탑 아이콘의 속성(Properties)의 Shortcut 탭의 Target 필드에 `-sdi`의 추가 등이 있다. Rcmdr의 제약은 Rcmdr 패키지가 의존하는 tcltk 패키지로부터 기인하는 것이다.

²Rcmdr의 실행은 R과 함께 배포되는 몇몇의 "추천(recommended)" 패키지들과 더불어 일부의 패키지를 요구함을 유의하자. Rcmdr, 요구되는 패키지들, 그리고 많은 기여 패키지들은 <<http://cran.r-project.org>>의 R 아카이브 네트워크(CRAN)에서 내려받을 수 있다.

만약 당신이 Rcmdr 패키지를 윈도우즈의 "R GUI"를 통해서 설치하면, Rcmdr이 의존하는 모든 패키지가 자동으로 드되지 않을 것이다. 만약 이러한 패키지들이 설치되지 않았다면, Rcmdr은 인터넷이나 컴퓨터 저장장치(예를 들어 CD/DVD)로부터 설치할 것을 제안할 것이다. 이것이 Rcmdr 패키지를 설치하는데 추천되는 절차이다. 대안적으로, Rcmdr와 의존성 패키지들을 `install.packages` 함수를 통해서 설치할 수 있다. 이 때 `dependencies=TRUE` 라는 인자를 설정해야 한다; 하지만, 이 방식은 당신이 정말로 원하는 패키지들보다 훨씬 많은 것들을 설치한다.

Dirk Eddebuettel의 기여로, 데비안 리눅스(Debian Linux) 사용자는 `$ apt-get install r-cran-rcmdr`의 명령어만으로 Rcmdr과 요구하는 의존성 패키지를 설치할 수 있다. 리눅스 시스템에서 Rcmdr 패키지를 빌드하고 설치하는 것은 바로 손쉽게 진행된다. 애플 맥(Mac)의 OS X의 경우는, Rcmdr이 의존하는 tcltk 패키지의 경우 X-Windows 용 Tcl/Tk의 사전 설치를 요구하고, R이 X-Windows 환경 아래에서 작동하기 때문에, 이 작업이 보다 일상천리로 이루어진다.

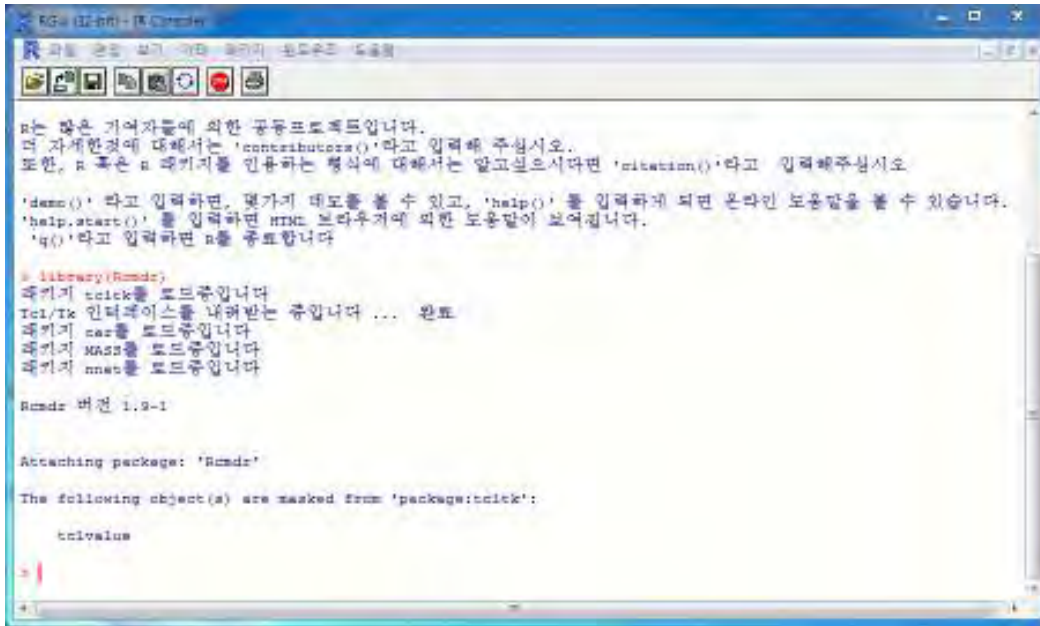


Figure 1: Rcmdr 패키지를 불러온 뒤의 R 콘솔(Console)



Figure 2: R Commander 시작 창.

- | - 결측치 사례 제거하기
- | - 활성 데이터셋 저장하기
- | - 활성 데이터셋 내보내기
- | 활성 데이터셋의 변수 관리하기 - 변수 다시코딩하기
 - | - 새로운 변수 계산하기
 - | - 데이터셋에 관측치 더하기
 - | - 변수 표준화하기
 - | - 수치형 변수를 요인변수로 바꾸기
 - | - 수치형 변수를 구간화하기
 - | - 요인 수준을 재정렬하기
 - | - 요인 대조 정의하기
 - | - 변수 이름 바꾸기
 - | - 데이터셋에서 변수 지우기
- 통계(Statistics) - 요약 - 활성 데이터셋
 - | - 수치적 요약
 - | - 빈도 분포
 - | - 관찰 결측치 계산하기
 - | - 통계표
 - | - 상관 행렬
 - | - 상관 검정
 - | - Shapiro-Wilk 정규성 검정
 - | - 분할표 - 이원표
 - | - 다원표
 - | - 이원표 입력 및 분석하기
 - | - 평균 - 일-표본 t-검정
 - | - 독립 표본 t-검정
 - | - Paired t-검정
 - | - 일원 분산 분석
 - | - 다원 분산 분석
 - | - 비율 - 일-표본 비율 검정
 - | - 이-표본 비율 검정
 - | - 분산 - 이-분산 F-검정
 - | - Bartlett의 검정
 - | - Levene의 검정
 - | - 비모수 검정 - 이-표본 Wilcoxon 검정
 - | - Paired-samples Wilcoxon 검정
 - | - Kruskal-Wallis 검정
 - | - Friedman 순위-합 검정
 - | - 척도 분석 - 척도 신뢰성
 - | - 주-성분 분석
 - | - 요인 분석
 - | - 확인적 요인 분석
 - | - 군집분석 - k-평균 군집 분석
 - | - 위계 군집 분석
 - | - 위계 군집화 요약하기
 - | - 위계 군집화를 데이터셋에 추가하기
 - | - 적합성 모델 - 선형 회귀
 - | - 선형 모델
 - | - 일반화 선형 모델
 - | - 다항 로짓 모델
 - | - 서열(Ordinal) 회귀 모델
- 그래프 - 색 팔레트
 - | - 색인 그림
 - | - 히스토그램
 - | - 줄기-잎 표시
 - | - 상자그림
 - | - 분위수-비교 그림
 - | - 산점도
 - | - 산점도 행렬
 - | - 선 그래프
 - | - XY conditioning plot

- | - 평균 그림
- | - Strip chart
- | - 막대 그래프
- | - 원 그래프
- | - 3차원 그래프 - 3차원 산점도
 - | - 마우스로 관측치 확인하기
 - | - 그래프를 파일로 저장하기
- | - 그래프를 파일로 저장하기 - bitmap(으)로
 - | - PDF/Postscript/EPS(으)로
 - | - 3차원 RGL 그래프
- 모델 - 활성 모델 선택하기
 - | - 모델 요약하기
 - | - 관찰 통계를 데이터에 추가하기
 - | - 신뢰 구간
 - | - Akaike 정보량 기준 (AIC)
 - | - Bayesian 정보량 기준 (BIC)
 - | - 단계적 모델 선택
 - | - 아위셋 모델 선택
 - | - 가설 검정 - 분산분석표
 - | - 두 모델 비교하기
 - | - 선형 가설
 - | - 수치적 진단 - 분산-팽창 요인
 - | - Breusch-Pagan 이분산성 검정
 - | - Durbin-Watson 자기상관 검정
 - | - 비선형성에 대한 RESET 검정
 - | - Bonferroni 이상치 검정
 - | - 그래프 - 기초 진단 그림
 - | - 잔차 분위수-비교 그림
 - | - 성분+잔차 그림
 - | - 추가-변수 그림
 - | - 영향 그림
 - | - 효과 그림
- 분포도 - 연속 분포 - 정규 분포 - 정규 분위수
 - | - 정규 확률
 - | - 정규 분포 그리기
 - | - 정규 분포의 표본
- | - t 분포 - t 분위수
 - | - t 확률
 - | - t 분포 그리기
 - | - t 분포의 표본
- | - 카이-제곱 분포 - 카이-제곱 분위수
 - | - 카이-제곱 확률
 - | - 카이-제곱 분포 그리기
 - | - 카이-제곱 분포의 표본
- | - F 분포 - F 분위수
 - | - F 확률
 - | - F 분포 그리기
 - | - F 분포의 표본
- | - 지수 분포 - 지수 분위수
 - | - 지수 확률
 - | - 지수 분포 그리기
 - | - 지수 분포의 표본
- | - 균일 분포 - 균일 분위수
 - | - 균일 확률
 - | - 균일 분포 그리기
 - | - 균일 분포의 표본
- | - 베타 분포 - 베타 분위수
 - | - 베타 확률
 - | - 베타 분포 그리기
 - | - 베타 분포의 표본
- | - Cauchy 분포 - Cauchy 분위수

- 가장 왼쪽의 (얇은) 단추(Button)는 활성화 데이터셋의 이름을 보여준다. 초기에는 활성화 데이터셋이 없다. 이 단추(Button)를 누르면, 현재 메모리에 있는 데이터셋들 중에서 (만약에 하나 이상이 있다면) 선택을 할 수 있게 된다. **R Commander**에 있는 대부분의 메뉴와 대화상자는 활성화 데이터셋에 대한 참조사항들이다. (파일(*File*), 편집(*Edit*), 분포(*Distributions*) 메뉴는 예외 사항이다.)

- 두개의 단추(Button)는 데이터셋을 수정하기 위한 **R** 데이터 편집기 또는 데이터셋을 조사하는 보기(Viewer) 창을 여는데 사용한다. 데이터셋 보기(뷰어, Viewer)는 다른 동작들이 실행될 때도 열린 상태로 남을 수 있다.³

- (얇은) 단추(Button)는 활성화된 통계 모델의 이름을 가리킨다 - 선형 모델(예를 들어, 선형 회귀 모델), 일반화 선형 모델, 다항 로짓 모델, 서열 회귀 분석.⁴ 초기에는 활성화된 모델이 없다. 하나 이상의 모델이 메모리에 있다면, 단추(Button)를 눌러서 그것들 중에서 선택할 수 있다.

- 도구막대(toolbar) 바로 아래에 스크립트 창 (그렇게 이름붙여진, 위아래-좌우로 펼 수 있는 (scrollable) 커다란 텍스트 창이 있다. 앞서 언급한 것처럼, GUI에 의해서 만들어진 명령문은 이 창안으로 복사된다. 당신은 스크립트 창에 있는 텍스트를 편집할 수 있고, 심지어 **R** 명령문을 이 창에 직접 입력할 수도 있다. 실행하기(*Submit*) 단추, 스크립트 창 오른쪽 아래 (또는, 대신, “실행(run)”의 *Ctrl-r*), 를 누르는 것은 커서를 담고있는 줄(line)을 실행(또는 재실행)하게 한다. 만약 여러개의 줄이 선택되면 (예를 들어, 마우스 왼쪽을 누르고 끌면서 그것들을 선택하면), 실행하기(*submit*)를 누를 경우 선택된 모든 줄이 실행하게 된다. 스크립트 창에 들어온 명령문은 한 줄 이상으로 확장될 수 있으나, 모든 줄이 한꺼번에 실행되어야 한다. *Ctrl-a*의 키조합은 스크립트 창 안에 있는 모든 텍스트를 선택하고, *Ctrl-s*는 창의 내용을 저장할 대화 상자를 불러온다.
- 스크립트 창 아래에는 출력물을 위해서 위아래-좌우로 펼 수 있고, 편집할 수 있는 커다란 텍스트 창이 있다. 이 창 안에 명령문은 붉은색으로 메아리처럼 복사되고, , 출력결과는 짙은 푸른색으로 인쇄된다 (마치 *R Console* 처럼).
- 바닥에는 메시지를 위한 작은 회색의 텍스트 창이 있다. 오류(errors) 메시지는 붉은 텍스트로, 경고(warnings)는 초록색으로, 다른 종류의 메시지는 짙은 푸른색으로 전시된다. 오류와 경고는 또한 종(bell)을 울리는 음향을 함께 제공한다.

당신이 **Rcmdr** 패키지를 불러온 이후, 콘솔(*R Console*)을 최소화시킬 수 있다. *R Commander* 창 또한 일반적인 방법으로 최대화 또는 크기조정이 가능하다. **R Commander**의 크기가 조정되면, R 출력물의 넓이는 출력 창에 적합하게 자동으로 수정된다.

R Commander는 대단히 설정(변경) 가능하다: 나는 여기까지 기본 설정을 묘사해왔다. 설정의 변화는 도구(*Tools*) - 선택기능(*Options*) ... 메뉴, 또는 - 훨씬 확장적으로 - **R**의 설정 기능으로 가능하다.⁵ **Rcmdr** 도움말 파일에서 자세한 사항을 살펴보라.

2 자료(데이터) 입력

R Commander의 대부분의 절차는 활성화된 데이터셋이 있다는 것을 가정한다.⁶ 만약에 메모리에 여러 개의 데이터셋이 있다면, 당신은 그 중 하나를 선택할 수 있으나, 오직 하나만 활성화된다. **R Commander**가 처음 시작할 때는 활성화된 데이터셋이 없다.

R Commander는 **R** 안으로 자료를 가져오는 여러 방법을 제공한다:⁷

- 데이터(*Data*) → 새로운 데이터셋...(New data set...) 을 통해서 자료(데이터)를 직접 입력할 수 있다. 매우 작은 데이터셋의 경우 합리적인 선택이다.
- 평범한-텍스트(“아스키(ascii)”) 파일 또는 클립보드로부터, 또다른 통계 패키지(**Minitab**, **SPSS**, 또는 **Stata**) 로부터, 또는 **Excel**, **Access**, 또는 **dBase** 으로부터 데이터셋을 불러들일 수 있다.

³David Firth의 relimp 패키지에 있는 `showData` 함수에 의해서 실행되는 데이터 뷰어는 많은 변수를 가진 데이터셋을 처리하는 데는 다소 느릴 수 있다. 만약 100개로 초기 설정된 반응치를 초과하는 숫자의 변수라면, 데이터셋을 전시하는데 다소 미학적으로 덜 기쁘게하는 **R View** 명령이 대신 사용된다. 변수의 숫자와 상관없이 **View** 를 사용하려면, 반응치를 0로 설정하라. 자세한 내용은 **R Commander**의 도움말 파일에서 찾아볼 것.

⁴**R Commander** plug-in 패키지들이 추가적인 모델 클래스를 제공할 것이다 (Fox, 2007; Fox and Carvalho, 2012).

⁵축약(예를 들어, 세계의 점, ...)으로 끝나는 메뉴 항목은 대화 상자를 이끈다; 이것은 표준화된 GUI 관례이다. 이 문서에서, → 는 어떤 메뉴 항목 또는 메뉴의 하위메뉴를 선택하는 것을 뜻한다.

⁶분포(*Distributions*) 메뉴를 통해서 선택된 절차들은 예외다. 예를 들어 통계(*Statistics*) → 분할표(contingency tables) 메뉴 아래에 있는 이원표 입력 및 분석하기(Enter and analyze two-way table)... 처럼.

⁷이 데이터 자원들 모두가 모든 플랫폼(운영체제)에서 이용가능하지는 않을 것이다.

- (만약 이름을 알고 있다면) 데이터셋의 이름을 입력함으로써, 또는 대화 상자에서 그 데이터셋을 선택함으로써 **R**의 패키지에 포함된 데이터셋을 읽을 수 있다.

2.1 텍스트 파일로부터 자료(데이터) 읽기

예를 들어, Nation.txt 라는 데이터 파일을 고려해보자.⁸ 이 파일 처음의 몇 줄은 다음과 같다:

	TFR	contraception	infant.mortality	GDP	region
Afghanistan	6.90	NA	154	2848	Asia
Albania	2.60	NA	32	863	Europe
Algeria	3.81	52	44	1531	Africa
American-Samoa	NA	NA	11	NA	Oceania
Andorra	NA	NA	NA	NA	Europe
Angola	6.69	NA	124	355	Africa
Antigua	NA	53	24	6966	Americas
Argentina	2.62	NA	22	8055	Americas
Armenia	1.70	22	25	354	Europe
Australia	1.89	76	6	20046	Oceania

...

- 파일의 첫 줄은 변수 이름을 담고 있다: TFR (the total fertility rate, 여성 한명 기준 자녀숫자 표현), contraception (기혼 여성의 피임 사용비율, %), infant.mortality (영아사망율, 1천명 출생 기준), GDP (1인당 국내 총생산, 미국 달러(\$)), 그리고 region(대륙이름).
- 그 아래 줄들은 각 국가에 한 줄씩 데이터 값 자체를 담고 있다. 데이터 값들은 "빈 공간(white space)" - 하나 이상의 공백 또는 탭 - 으로 구분된다. 데이터 값을 수직으로 배열하는 것이 좋지만, 반드시 그렇게 하는 것이 필요하지는 않다. 데이터의 줄은 국가 이름으로 시작됨을 주시하라. 왜냐하면, 우리는 이러한 국가 이름들이 데이터셋의 "열 이름(row names)" 이 되기를 바라기 때문에, 상응하는 변수이름이 없다: 다시 말하면, 다섯 개의 변수 이름이 있지만, 각 줄에는 여섯 개의 데이터 값이 있다. 이러한 상황이 발생하면, **R**은 각 줄의 첫 번째 값을 열 이름으로 해석할 것이다.
- 일부 데이터 값이 보이지 않는다. **R**에서 결측 데이터를 코딩하는데, 내가 여기서 한 방식처럼, NA("not available"(결측)을 대변하는) 를 사용하는 것이 가장 편리하다.
- 변수 TFR, contraception, infant.mortality, 그리고 GDP 는 수치형 (양적) 변수들이다; 대조적으로, region 은 지역 이름을 담고 있다. 데이터를 읽을 때, **R**은 region 을 "요인(factor)" - 다른 말로, 범주형 변수 (categorical data) - 로 취급한다. 대부분의 맥락에서, **R Commander** 는 수치형 변수와 요인을 구별한다.

R 안으로 데이터를 읽어 들이려면, *R Commander* 메뉴에서 데이터 (Data) → 데이터 불러오기 (Import data) → 텍스트 파일, 클립보드, 또는 URL에서...(from text file, clipboard, or URL...) 을 선택하라. 이 동작은 그림 3. 에서 보이는 것처럼, 텍스트 데이터 읽기 대화상자를 화면에 띄운다. 기본 설정된 데이터셋의 이름은 데이터셋(Dataset) 이다. 나는 이름을 Nations로 바꾼 것이다.

유효한 **R** 이름은 (영어) 대문자 또는 소문자 (또는 마침표, .)로 시작하고, 문자, 마침표, 아랫줄(underscores, _), 숫자(예, 0-9)로 구성된다; 특히 주의할 사항으로, 데이터셋 이름 중간에 공백을 포함시키지 마라. 또한 **R**은 대-소문자 구별, 예를 들어 nations, Nations, 그리고 NATIONS 등이 구별되며, 다른 데이터셋을 표현하는데 사용될 수 있다.

텍스트 데이터 읽기(Read Text Data) 대화 상자의 예(OK) 단추를 누르면, 그림 4. 처럼 열기(Open) 파일 대화상자가 나타난다. 여기서 나는 Nations.txt 를 찾아 이용했다. 대화 상자의 열기(Open) 단추를 누르면, 읽혀질 데이터 파일 목록을 보게될 것이다. 데이터 파일이 읽혀지면, 이것은 곧 **R Commander**의 활성화 데이터셋이 된다. 결과적으로, 그림 5. 처럼 데이터셋의 이름은 *R Commander* 창의 왼쪽 상단 근처의 데이터셋 단추에 나타나게 된다.

나는 또한 그림 5. 에서 보이는 것처럼, 데이터셋을 불러오기 위해서 데이터셋 보기(View data set) 단추를 눌렀다. Nations 라는 데이터셋을 읽고 보기 위한 명령 (**R**의 read.table과 showData 명령) 이 데이터셋의 등장에 의해서 부분적으로 가려지면서, 스크립트 창과 출력 창에 나타난다. 데이터셋이 읽혀지고, 활성화 데이터셋이 되면, 메시지 창에 주석 하나가 나타난다.



Figure 3: 텍스트 파일로부터 데이터 읽기.

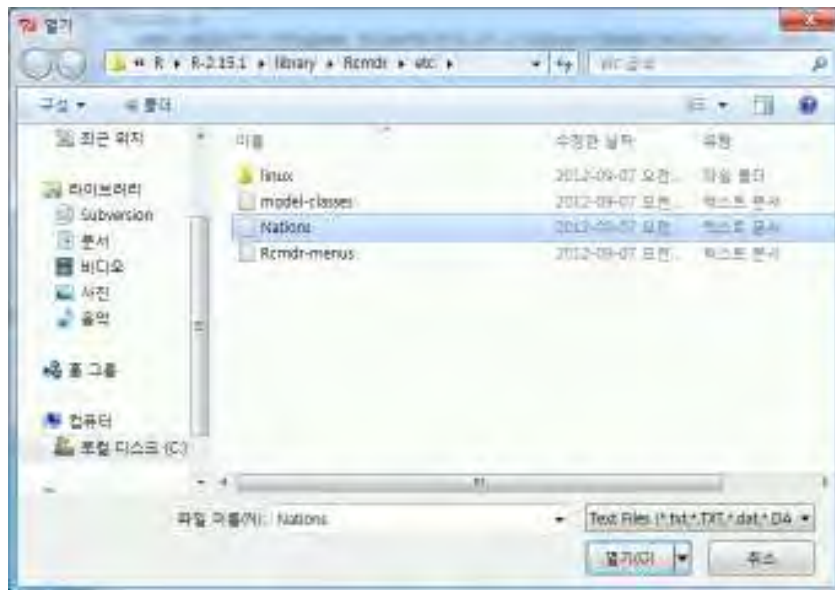


Figure 4: 텍스트 파일을 읽기위한 파일-열기 대화상자.

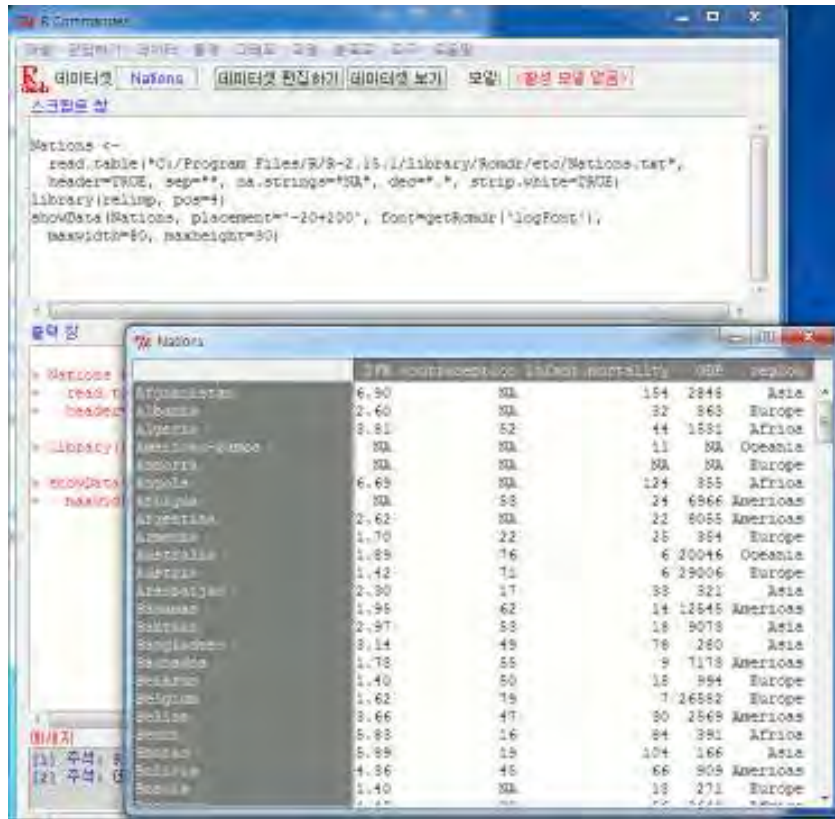


Figure 5: 활성 데이터셋 배열하기.

`read.table` 명령은 R "데이터 프레임 (data frame)", 사각형의 사례-변수 형식의 데이터셋을 담고 있는 객체, 를 생성한다. 데이터셋의 열은 사례 또는 관찰(관측치)을, 행은 변수를 표현한다. R Commander의 데이터셋은 R 데이터 프레임이다.

2.2 데이터 직접 입력하기

스프레드시트(spreadsheet, 전표) 같은 R의 기초적인 데이터 편집기에 데이터를 직접 입력하기 위해서 아래와 같은 절차를 밟을 수 있다: 예를 들면, Moore(2000)에 있는 Problem 2.44로부터 아주 작은 데이터셋을 사용한다:

- R Commander 메뉴의 데이터(Data) → 새로운 데이터셋...(New data set...)을 선택한다. 선택적으로 데이터셋의 이름을, 예를 들어 Problem2.44, 대화 상자에 입력하고, 예(OK) 단추를 누른다. (R 이름은 공백을 포함시킬 수 없음을 기억하라.) 이것은 빈 데이터셋을 가진 데이터 편집기(Data Editor)를 띄울 것이다.
- 데이터 편집기의 첫 두개의 행(columns) 안에 the problem의 데이터를 입력한다. 자판의 화살표 키를 사용함으로써 셀 하나씩을 이동할 수 있다. 탭 키를 이용하거나, 실행(enter) 키를 누르거나, 마우스로 위치를 가리키고 왼쪽 마우스를 누르거나 해도 된다. 데이터셋 입력이 끝나면, 창은 그림 6. 처럼 보여야 한다.
- 다음, 첫 행(column) 위에 있는 var1 이름을 누른다. 이것은 그림 7. 처럼 변수 편집기(Variable editor) 대화 상자를 띄울 것이다.
- 내가 했던 것처럼, 상자에 변수 이름 age 를 타자치고, 변수(variable editor) 편집기 창의 오른쪽 상단 구석에 있는 X 단추를 누르거나, 실행(enter) 키를 눌러서 창을 닫는다. 두 번째 행(column)에 height 라는 이름을 넣고 이 과정을 반복한다. 이제 데이터 편집기(Data Editor)가 그림 8. 처럼 보여야 한다.

⁸이 파일은 Rcmdr 패키지의 etc 하위디렉토리에 담겨 있다.

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6
1	30	50				
2	40	90				
3	51	91				
4	54	93				
5	57	94				
6	60	96				
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

Figure 6: 데이터 입력 이후의 데이터 편집기.

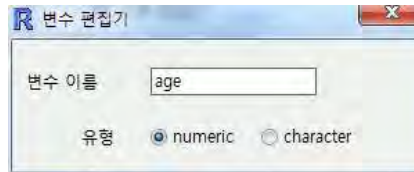


Figure 7: 데이터 편집기에서 변수 이름 바꾸는 대화 상자.

- 데이터 편집기(Data editor) 메뉴에서 파일(File) → 닫기(Close) 를 선택하거나 또는 데이터 편집기(Data editor) 창의 오른쪽 상단의 X 를 누른다. 당신이 입력한 데이터셋은 이제 R Commander의 활성화 데이터셋이 된다.

2.3 패키지에서 데이터 읽기

많은 R 패키지가 데이터를 포함하고 있다. 패키지에 있는 데이터셋의 목록은 데이터(Data) → 패키지에 있는 데이터 (Data in packages) → 패키지의 데이터셋 목록 (List data sets in packages) 을 통하여 팝업-창에 나타난다. 그리고 Data (데이터) → 패키지에 있는 데이터 (Data in packages) →첨부된 패키지에서 데이터셋 읽기 (Read data set from an attached package) 를 통하여 R Commander 안에서 읽을 수 있다.⁹ 결과 대화 상자는 그림 9. 와 같다. 만약 당신이 어느 패키지에 있는 데이터셋 이름을 알고 있다면, 그 이름을 직접 입력할 수 있다; 그렇지 않으면 패키지 이름을 두번 누름(double-clicking)으로 오른쪽의 목록 상자에 그 패키지에 포함된 데이터셋을 배열시킬 수 있다.¹⁰ 자판에서 어떤 문자키를 누르면 데이터셋의 목록이 해당 문자키로 시작되는 다른 데이터셋의 위치로 이동일 것이다. 도구(Tools) → 패키지 불러오기(Load packages) 를 통하여 추가적인 R 패키지들에 접근할 수 있다.

3 수치적 요약과 그래프 생성하기

활성화 데이터셋이 있다면, 다양한 수치적 요약과 그래프를 생산하기 위하여 R Commander 메뉴를 사용할 수 있다. 여기서 몇 개의 기초 예제를 묘사하고자 한다. 좋은 GUI는 크해보야 자기-설명적이어야 한다. 나는 당신이 어떻게 R Commander가 작동하는지를 보고 나면, 아마도

⁹ 패키지에 있는 모든 데이터셋이 데이터 프레임은 아니다. 오직 데이터 프레임만이 R Commander의 사용에 적절하다.

¹⁰ 일반적으로 R Commander에서, 어떤 목록 상에서 항목을 복사하여 대화상자의 다른 위치로 이동시키는 것이 필요할 때, 두번 누름(double-clicking)이 요구된다.

순번	변수명	변수값	변수명	변수값	변수명	변수값
1	36	36				
2	37	37				
3	38	38				
4	39	39				
5	40	40				
6	36	36				
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

Figure 8: 두 변수 이름이 바뀌어진 이후의 데이터 편집기.

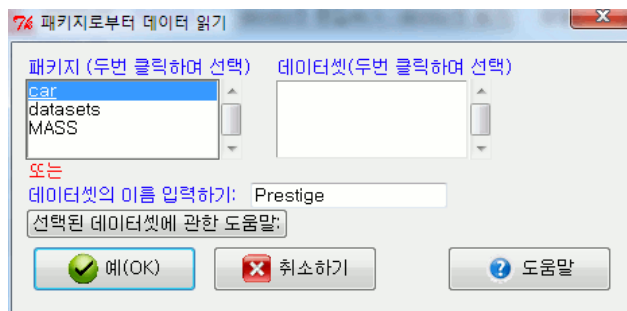


Figure 9: 첨부된 패키지에서 자료 읽기 - car 패키지의 Prestige 데이터셋의 경우.

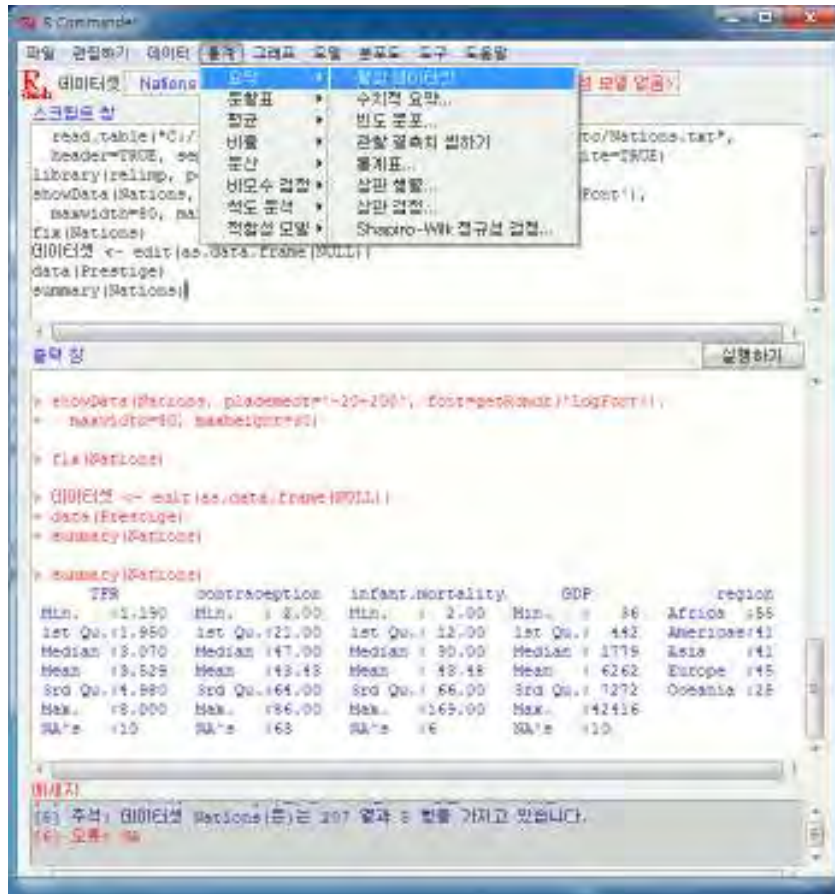


Figure 10: 활성화 데이터셋의 변수 요약 얻기.

온라인 도움말 파일들의 도움을 받아서, 당신이 그것을 사용하는 데 거의 어려움을 겪지 않기를 희망한다.

아래의 몇 예제에서, 앞 장에서 불러온 텍스트 파일에서 읽혀진 Nations (이)라는 데이터셋이, 활성화 데이터셋이라고 가정한다. 만약에 Moore(2000)의 다섯-관찰 데이터셋에서 타자 입력했거나, car 패키지의 Prestige 데이터셋을 읽어들었다면 - 또한 앞장에서 이미 묘사했던 동작들임 - 이것들 중의 하나가 활성화된 데이터셋이다. 기억하라. 당신은 현재 메모리 상주한 데이터셋들 중에서 하나를 선택할 때, R Commander 창의 왼쪽 상단 근처에 있는 활성화 데이터셋의 이름과 연관된 얇은 단추(flat button)를 눌러서 활성화 데이터셋을 변경할 수 있다는 사실을.

통계 (Statistics) → 요약 (Summaries) → 활성화 데이터셋 (Active data set) 을 선택하면 그림 10. 에 있는 결과를 생산한다. 데이터셋의 각 수치형 변수 (TFR, contraception, infant.mortality, 그리고 GDP) 에 대하여, R은 결측치와 더불어 최소값, 최대값, 첫째-세째 분위수, 중앙값, 평균 등을 보고한다. region 이라는 범주형 변수에 대하여, 요인의 각 '수준'에서 관찰 사례의 숫자를 가진다. 데이터셋이 열 개 이상의 변수를 가졌다면, R Commander는 당신이 다음 단계로의 진행을 진심으로 원하는지 확인을 요구 - 원하지 않는 대규모의 결과물을 생산하는 것으로부터 잠재적으로 우리를 방어하는 - 했을 것이다.

비슷하게, 통계(Statistics) → 요약(Summaries) → 수치적 요약... (Numerical summaries...) 의 선택은 그림 11. 과 같은 대화 상자를 띄울 것이다. 오직 수치형 변수만 이 대화 상자의 변수 목록에 보이게 된다; region 이라는 요인은 빠져있다. 왜냐하면, 요인을 위한 수치적 요약을 계산하는 것은 민감하지 않기 때문이다. infant.mortality 를 누르고, 예(OK)를 누르면, (출력창에) 다음의 출력결과를 생산한다.¹¹

¹¹ 변수-목록 상자에서 변수 한개를 선택하려면, 단지 그 이름을 마우스 왼쪽으로 누르면 된다. 다른 맥락에서, 당신은 한개 이상의 변수를 선택해야만 한다, 또는 선택하기 원할 수 있다. 이런 경우에는 일반적인 윈도우즈 (Windows) 관례가 적용된다: 마우스 왼쪽으로 변수하나를 선택하고, 이전에 선택된 다른 변수들을 선택해제 한다; Shift-left-click 은 선택을 확장한다; 그리고 Ctrl-left-click 으로 개별 선택을 추가한다.



Figure 11: 수치적 요약 대화 상자.

```
> numSummary(Nations["infant.mortality"], statistics=c("mean", "sd", "IQR",
+ "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

```
      mean      sd IQR 0% 25% 50% 75% 100%   n NA
43.47761 38.75604  54  2  12  30  66  169 201  6
```

기본 설정으로, 실행되는 R 명령은 변수의 평균과 표준편차 (sd), interquartile range(IQR) 와, 최소값, 첫째 분위수, 중앙값, 세째 분위수, 최대값에 상응하는 분위수들을 (백분위들) 출력한다; n 은 유효한 관찰 사례 숫자, NA 는 결측치의 숫자다.

R Commander 대화상자의 전형적인 모습으로, 그림 11.의 수치적 요약 (Numerical Summaries) 대화 상자는 예(OK), 취소하기(Cancel), 그리고 도움말(Help) 단추를 포함한다. 도움말(Help) 단추는 대화 상자 자체 또는 (여기처럼) 대화상자를 이끌어내는 R 함수의 도움말 페이지를 (웹 브라우저를 통해서) 불러낸다.¹²

수치적 요약(Numerical Summaries) 대화 상자는 또한 요인 수준에서 정의된 집단들의 요약을 계산하는 항목을 만든다. 집단별 요약...(Summarize by groups...) 단추는 그림 12. 처럼 집단(Groups) 대화상자를 이끌어 낸다. Nations 라는 데이터셋에는 오직 하나의 요인만 있기 때문에, 변수 목록에 region 변수만 보이고, 사전에 이미 선택된다; 예(OK)를 누르면 집단별 요약...(Summarize by groups...) 단추가 summarize by region 으로 바뀐다 (그림 13. 참조); 다시 예(OK)를 누르면 다음의 결과가 생산된다:

```
> numSummary(Nations[,c("GDP", "infant.mortality")], groups=Nations$region,
+ statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

Variable: GDP

```
      mean      sd      IQR 0%  25%  50%  75% 100%   n NA
Africa  1196.000 2089.614  795.50 36 209.00 389.5 1004.50 11854 54  1
Americas 5398.000 6083.311 5268.50 386 1749.25 2765.5 7017.75 26037 40  1
Asia    4505.051 6277.738 6062.50 122 345.00 1079.0 6407.50 22898 39  2
Europe  13698.909 13165.412 24582.25 271 1643.75 9222.5 26226.00 42416 44  1
Oceania  8732.600 11328.708 16409.25 654 1102.75 2348.5 17512.00 41718 20  5
```

Variable: infant.mortality

```
      mean      sd IQR 0%  25%  50%  75% 100%   n NA
Africa  85.27273 35.188095 50.0  7 61.00 85.0 111.00  169 55  0
Americas 25.60000 17.439713 24.0  6 12.00 21.5  36.00  82 40  1
Asia    45.65854 32.980001 50.0  5 22.00 37.0  72.00  154 41  0
Europe  11.85366  7.122363 10.0  5  6.00  8.0  16.00  32 41  4
Oceania 27.79167 29.622229 26.5  2  9.25 20.0  35.75  135 24  1
```

¹²대화상자는 재설정(Reset) 단추를 가질 수 있다: 3.1장의 대화-상자 메모리의 묘사를 참조하라.



Figure 12: 집단(Group) 대화 상자에서 집단화 변수 선택하기.



Figure 13: 두 개의 수치형 변수와 region 이라는 집단화 변수를 선택한 이후의 수치적 요약 (numerical summaries) 대화 상자.

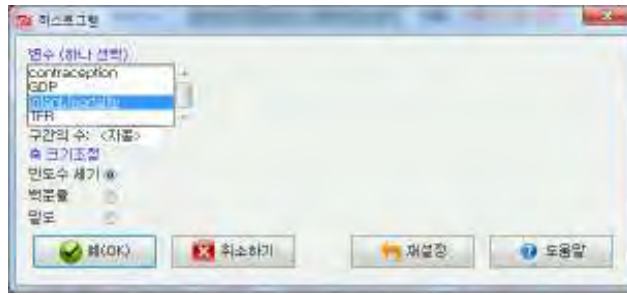


Figure 14: 히스토그램 대화상자.

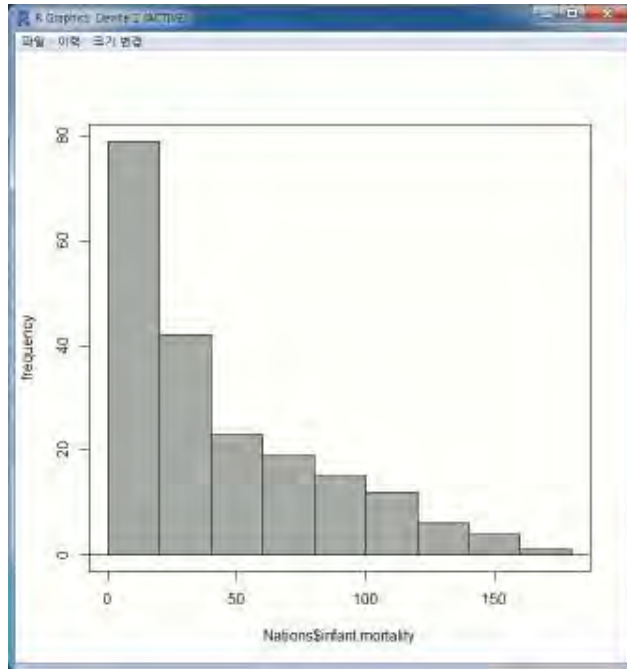


Figure 15: Nations 데이터셋의 infant mortality에 대한 히스토그램을 담고 있는 그래픽 창.

여러 개의 다른 **R Commander** 대화상자를 통해서 이 방식으로 집단 변수를 선택할 수 있다.

R Commander로 그래프를 만드는 것도 또한 곧바로 된다. 예를 들어, *R Commander* 메뉴의 **그래프(Graphs)** → **히스토그램...(Histogram...)** 선택은 그림 14. 처럼 히스토그램(*Histogram*) 대화상자를 띄운다; 그리고 *infant.mortality* 를 선택한 후 **예(OK)**를 누르면, 그림 15. 처럼 히스토그램의 그래픽 장치 창이 열린다.

만약 세션 하나에 여러 개의 그래프를 만든다면, 그래픽 장치 창에 일반적으로 오직 가장 최근의 그래프가 보인다. 자판의 *Page Up* 과 *Page Down* 키를 사용하면서, 이전의 그래프를 불러낼 수 있다.¹³

4 통계 모델

R Commnader의 **통계(Statistics)** → **적합성 모델(Fit Models)** 메뉴 아래에서 여러 종류의 통계 모델이 이용가능하다: (**선형 회귀(Linear regression)**)와 **선형 모델(Linear model)** 양자를 바

¹³시작할 때, **R Commander**는 그래프 기록(history) 메카니즘을 켜다; 이 특징은 오직 윈도우즈(**Windows**) 시스템에서 이용 가능하다. **그래프(Graphs)** → **3차원 그래프(3D graph)** → **3차원 산점도...(3D scatterplot...)** 를 통하여 생성되는 역동적인 3차원 산점도는 특별히 *RGL* 장치에서 나타난다; 마찬가지로 통계 모델을 위해서 생성되는 효과 배열(effect displays)은 (Fox, 2003; Fox and Hong, 2009)은 모델(*Models*) → **그래프(Graphs)** → **효과 그림(Effect plots)** 을 통하여 개별적인 그래픽-장치 윈도우즈에 나타난다.



Figure 16: 선형 모델(Linear Model) 대화 상자 (car 패키지의 Prestige가 활성 데이터셋이 된 상태)

탕으로) 선형 모델(Linear models), 일반화 선형 모델 (Generalized linear models), 다항 로짓 모델(multinomial logit models), 그리고 proportional-odds 모델과 같은 서열 회귀 모델(ordinal regression models) [마지막 두개는 각각 Venables와 Ripley (2002)의 **nnet**, **MASS** 패키지의 함수 사용]. 비록 세부적인 부분에서는 결과 대화 상자들이 차이가 있을지라도, 예를 들어 일반화 선형 모델 대화상자는 분포류(distributional family)와 대응연결 함수를 선택하는데 준비를 제공한다, 그것들은 그림 16. 에서 선형 모델 (Linear Model) 대화방식이 예시되듯이, 공통의 일반 구조를 공유하고 있다.¹⁴

- 변수-목록 상자의 변수 하나를 두번 클릭하면, 그것은 모델 공식에 복사된다 - 만약 비어있다면, 공식의 왼쪽에, 그렇지 않으면 오른쪽에 (맥락이 요구하는 상황에서 + 부호를 먼저 입력하면서). 요인(범주형 변수)은 변수 목록에 괄호 형태로 꼬리표가 붙는다.
- 위 공식의 단추 옆은 공식의 오른쪽에 입력될 수 있는 연산자와 괄호로 사용될 수 있다.
- 또한 공식 공간 안에 직접 입력이 가능하다. 예를 들어, 공식에 $\log(\text{income})$ 과 같은 용어를 입력하려면 직접 입력해야만 한다.
- 모델의 이름은, 여기서는 `LinearModel.1`, 자동으로서 산출된다. 하지만, 유효한 어떤 **R** 이름으로도 대체가 가능하다.
- 하위셋 표현식(Subset expression)이라고 이름이 붙은 상자에 **R**의 표현식을 타자로 입력할 수 있다; 만약 내용이 공급되면, 이것은 회귀 함수 (lm function)의 하위 인자(subset argument)로 통과된다, 그리고 데이터셋 관측치의 하위셋의 모델 적합성에 사용된다. 하위셋 표현식의 한 형태는 각 관측치의 참(TRUE) 또는 거짓(FALSE)을 평가하는 논리적 표현이다: 예를 들어 (Prestige 데이터셋에서 모든 비전문가적 직업을 선택하게 되는) `type != "prof"`.

예(OK) 단추를 누르면 (출력창에) 다음의 결과물을 생산한다. 그리고 모델(Model) 단추에 `LinearModel.1`이라는 이름을 전시하며, `LinearModel.1`을 활성 모델로 만든다.

```
> LinearModel.1 <- lm(prestige ~ (education + income)*type, data=Prestige)
> summary(LinearModel.1)
```

Call:

```
lm(formula = prestige ~ (education + income) * type, data = Prestige)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-13.462  -4.225   1.346   3.826  19.631
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
```

¹⁴최근의 데이터셋의 수치형 변수를 담고 있는 목록 상자에서 단순히 반응변수와 설명변수의 이름을 선택하는 선형 회귀(Linear Regression) 대화는 예외가 된다. 아래의 설명은 R 모델 공식과의 친숙함을 가정한다; 예를 들어, R 콘솔(Console)의 도움말(Help) 메뉴로부터 접근될 수 있는, R과 함께 오는 R 소개(Introduction to R) 사용설명서를 참조하라.

```

(Intercept)      2.276e+00  7.057e+00  0.323  0.7478
education        1.713e+00  9.572e-01  1.790  0.0769 .
income           3.522e-03  5.563e-04  6.332  9.62e-09 ***
typeprof         1.535e+01  1.372e+01  1.119  0.2660
typepcw         -3.354e+01  1.765e+01 -1.900  0.0607 .
education:typeprof 1.388e+00  1.289e+00  1.077  0.2844
education:typepcw  4.291e+00  1.757e+00  2.442  0.0166 *
income:typeprof  -2.903e-03  5.989e-04 -4.847  5.28e-06 ***
income:typepcw   -2.072e-03  8.940e-04 -2.318  0.0228 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 6.318 on 89 degrees of freedom
(4 observations deleted due to missingness)
Multiple R-squared:  0.8747,    Adjusted R-squared:  0.8634
F-statistic: 77.64 on 8 and 89 DF,  p-value: < 2.2e-16

```

모델(*Models*) 메뉴로부터 활성 모델에 관한 조작기능을 선택할 수 있다. 예를 들어, 모델(*Models*)→가설 검정(*Hypothesis tests*)→ 분산분석표(*Anova table*)는, 기본 설정인 “Type-II”의 선택 상황에서, 다음의 결과물을 생산한다:

```

> Anova(LinearModel.1, type="II")

Anova Table (Type II tests)

Response: prestige
      Sum Sq Df F value    Pr(>F)
education 1068.0  1 26.7532 1.413e-06 ***
income    1131.9  1 28.3544 7.511e-07 ***
type       591.2  2  7.4044  0.00106 **
education:type 238.4  2  2.9859  0.05557 .
income:type  951.8  2 11.9210 2.588e-05 ***
Residuals 3552.9 89
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

5 추가설명(Odds and Ends)

5.1 출력결 저장 및 인쇄하기

텍스트 출력물을 *R Commander*의 파일(File)에서 직접 저장할 수 있다. 마찬가지로 *R Graphics* 장치 창의 파일(File) 메뉴에서 그래프를 저장하거나 인쇄할 수 있다. 하지만, 일반적으로 보관하고자 하는 텍스트 출력물과 그래프를 워드-프로세스 문서 안에 수집하는 것이 더 편리하다. 이러한 방법으로 R 결과물에 타자로 입력한 주석과 설명을 덧붙일 수 있다.

Word, OpenOffice Writer 또는 심지어 **Windows Wordpad** 등의 워드 프로세스를 열 어라. 출력 창에서 텍스트를 복사하려면, 마우스로 텍스트를 블록화하여 선택하고, 편집하기(*Edit*) 메뉴에서 복사하기(*Copy*)를 선택하고 (또는 *Ctrl-c* 라는 키조합을 함께 누르거나, 창에서 마우스 오른쪽 단추를 눌러서 맥락 메뉴에서 복사하기(*Copy*)를 선택하고), 윈도우즈 프로그램(applications)에서 편집(*Edit*) → 붙이기(*Paste* (또는 *Ctrl-v*))를 통하여 워드-프로세서 안으로 텍스트를 붙인다. 도움말 하나는 제시한다: **R**의 텍스트 출력에 *Courier New*와 같은 **mono-spaced**("typewriter"■) 폰트를 사용하라, 그렇지 않으면 결과물이 이쁘게 줄맞추어 정돈되지 않는다.

마찬가지로, 그래프를 복사하려면, R 그래픽 장치 메뉴에서 파일(File) → 클립보드에 복사하기(*Copy to the clipboard*) → *as a Metafile*을 선택하라; 그리고 그래프를 편집하기(*Edit*) → 붙이기 (또는 *Ctrl-v*)를 통하여 워드-프로세서 안에 그래프를 붙여라. 대안적으로, **R** 그래픽 장치에서 그래프를 복사하기 위하여 *Ctrl-w*, 또는 맥락 메뉴를 불러오기 위하여 그래프 위에서 마우스 오른쪽을 클릭할 수 있다. 이 방식으로 *Copy as metafile*을 선택할 수 있다.¹⁵ R 세션의 마지막에서 작업 기록의 주석을 제공하는, 당신이 생성한 문서를 저장 또는 인쇄할 수 있다.

¹⁵ 이러한 메뉴들을 조사할 때 보게 되듯이, 여러 가지 다양한 형식으로 그래프를 메모리(클립보드)에 저장하는 것 뿐만 아니라 파일로도 저장할 수 있다. 하지만, 여기서 제안되는 절차는 곧바로 그리고 일반적으로 고-품질 그래프 결과물 나타낸다. 다시 한번 말하건대, 이 묘사는 윈도우즈 시스템에 해당된다.

다른 대안으로, *R Commander*의 파일(*File*)과 그래프(*Graphs*) → 그래프를 파일로 저장하기 (*Save graph to file*) 메뉴에서 텍스트와 그래프 출력결과를 저장할 수 있다. 파일(*File*) → 스크립트 저장하기(*Save script*)를 통하여, **R Commander** 스크립트 창을 저장하면, 나중에 기존 작업을 재생산할 수 있다.

5.2 R 세션 종료하기

세션을 종료하는 여러가지의 방법이 있다. 예를 들어, *R Commander* 메뉴의 파일(*File*) → 끝내기(*Exit*) → *Commander*와 *R*에서(*from Commander and R*)를 선택할 수 있다. 종료 확인을 질문받을 것이며, 스크립트 창과 출력 창의 내용을 저장하기 원하는지 질문받을 것이다. 마찬가지로, *R 콘솔(Console)*에서 파일(*File*) → 끝내기(*Exit*)를 선택할 수 있다; 이 경우, (**R**이 메모리에 간직하고 있는 데이터셋과 같은) **R** 작업공간을 저장하기 원하는지 질문을 받을 것이다; 일반적으로 아니오(*No*)라고 대답할 것이다.

5.3 스크립트 창에서 명령 입력하기

스크립트 창은 편집, 입력과 명령어 실행을 위한 간단한 편의도구를 제공한다. **R Commander**에 의해서 생성된 명령어는 스크립트 창에 나타나며, 다른 편집기처럼 명령어를 타자치거나 편집할 수 있다. 하지만, **R Commander**는 진정한 의미의 **R 콘솔(Console)**을 제공하는 것은 아니며, 스크립트 창은 일부 한계를 갖고 있다. 예를 들어, 여러 줄로 된 명령어의 모든 줄은 실행을 위하여 동시에 한꺼번에 제출(실행)되어야만 한다. 복잡한 **R** 프로그래밍을 위해서는 윈도우즈(**Windows**)와 **Mac OS X** 버전의 **R**에서 제공되는 스크립트 편집기, 또는 - 나아가 - 프로그래밍 편집기, 또는 **RStudio** <www.rstudio.org>와 같은 대화형 개발 환경(interactive development environment, IDE)를 선택할 수 있다.¹⁶

References

- [1] Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15):1–27.
- [2] Fox, J. (2005). The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software*, 19(9):1–42.
- [3] Fox, J. (2007). Extending the R Commander by “plug in” packages. *R News*, 7(3):46–52.
- [4] Fox, J. and Carvalho, Marilia S. (2012). The RcmdrPlugin.survival package: Extending the R Commander interface to survival analysis. *Journal of Statistical Software*, 49(7):1–32.
- [5] Fox, J. and Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, 32(1):1–24.
- [6] Moore, D. S. (2000). *The Basic Practice of Statistics*, Second Edition. Freeman, New York.
- [7] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth Edition. Springer, New York.

¹⁶**R Commander**는 **RStudio** 환경 아래서 작동할 것이다. 이 경우, 기본 설정으로 **R Commander**의 결과물과 전달글(messages)은 **RStudio** 안의 **R console**로 바로 향하게 된다. 하지만, **RStudio**의 그래픽 장치와의 연동에서 불안정성과 비호환성 등의 몇몇 쟁점이 있다.