

R 소개

R 프로그램의 History

- R 프로그램은 미국 벨(Bell) 연구소의 John Chambers가 개발한 S언어를 기반으로 만들어짐
- 1993년 뉴질랜드의 오클랜드대학교(University of Auckland)의 로스 이하카(Ross Ihaka)와 로버트 젠틀맨(Robert Gentleman)에 의해서 개발
- SPSS, STATA, SAS등의 유사한 소프트웨어들이 있고, 요즘에는 Python도 사용함



R 프로그램의 History

Release	Date	Description
0.16		This is the last alpha version developed primarily by Ihaka and Gentleman. Much of the basic functionality from the "White Book" (see S history) was implemented. The mailing lists commenced on April 1, 1997.
0.49	1997-04-23	This is the oldest source release which is currently available on CRAN. ^[46] CRAN is started on this date, with 3 mirrors that initially hosted 12 packages. ^[47] Alpha versions of R for Microsoft Windows and the classic Mac OS are made available shortly after this version. ^[citation needed]
0.60	1997-12-05	R becomes an official part of the GNU Project . The code is hosted and maintained on CVS .
0.65.1	1999-10-07	First versions of <code>update.packages</code> and <code>install.packages</code> functions for downloading and installing packages from CRAN. ^[48]
1.0	2000-02-29	Considered by its developers stable enough for production use. ^[49]
1.4	2001-12-19	S4 methods are introduced and the first version for Mac OS X is made available soon after.
2.0	2004-10-04	Introduced lazy loading , which enables fast loading of data with minimal expense of system memory.
2.1	2005-04-18	Support for UTF-8 encoding, and the beginnings of internationalization and localization for different languages.
2.11	2010-04-22	Support for Windows 64 bit systems.
2.13	2011-04-14	Adding a new compiler function that allows speeding up functions by converting them to byte-code.
2.14	2011-10-31	Added mandatory namespaces for packages. Added a new parallel package.
2.15	2012-03-30	New load balancing functions. Improved serialization speed for long vectors.
3.0	2013-04-03	Support for numeric index values 2^{31} and larger on 64 bit systems.
3.4	2017-04-21	Just-in-time compilation (JIT) of functions and loops to byte-code enabled by default.
3.5	2018-04-23	Packages byte-compiled on installation by default. Compact internal representation of integer sequences. Added a new serialization format to support compact internal representations.

주요 특징

- Open Source
 - 오픈 소스로 개인, 기관, 기업에서 무료로 사용 가능
- 데이터 분석
 - 다양한 통계 방법론을 적용한 데이터 분석 기능을 제공 R 프로그램은 데이터 분석을 목적으로 만들어졌기 때문에 데이터 분석에 필요함
- 강력한 그래프 기능
 - 2D, 3D 그래픽, 지도, GIS, 동적 그래프 등을 지원
- 데이터 핸들링
 - 텍스트, CSV, 엑셀, SAS, SPSS, Stata, DB 등의 다양한 데이터를 읽어오는 기능 수정, 삭제, 정렬, 합치기 등의 데이터 핸들링을 위한 기능
- 메모리
 - 데이터는 메모리(RAM)에서 작동되기 때문에 데이터 처리가 빠름 ※메모리 크기에 따라 분석할 수 있는 데이터의 양이 결정됨

장점

- GPL로 오픈 소스로 배포되고 있어 무료로 사용할 수 있다. SPSS, MATLAB과 같은 상용 프로그램을 구입하지 않아도 됨.
- R에서 사용할 수 있는 수많은 통계 관련 패키지가 개발되어 있어서 인터넷을 통해 이 패키지들을 설치하는 식으로 무수한 기능 확장이 가함.
- 통계학자들이 만들어 낸 언어이며 통계 전문 언어 중 가장 보편적이기 때문에 내가 사용하고 싶은 모든 통계 기법이 이미 어딘가에 패키지 형태로 구현되어 있다고 봐도 됨.
- 그래픽 관련 패키지를 설치하면 간단하게 다양한 그래프를 활용할 수 있으며 구글이나 네이버 지도를 불러오거나 이를 활용해 GIS 용도로 쓰는 것도 가능함.
- 웹 어플리케이션 개발 프레임워크인 Shiny의 고도화로 통계 또는 머신러닝 모델을 웹과 연동할 수 있음.
- 데이터 마이닝, 빅 데이터 프로세싱, 기계학습 등에 유용함.
- 리스크, 재무, 마케팅 담당자 채용시 R 능통자를 우대하기도함.

단점

- 메모리
 - 큰 데이터 집합을 이용할 때 문제가 발생할 수 있음.
 - 데이터를 물리적 메모리에 저장해야 하기 때문에 효율이 중시되는 프로젝트에서는 먼저 R로 구현한 후 그걸 C 등의 일반 프로그래밍 언어로 포팅하는 경우가 많음.
 - 다른 경우는 복잡한 데이터 작업은 C 또는 Fortran(포트란)에서 작업시키고 결과만 가져오는 방식을 사용하기도 함. 다만, 컴퓨터에 장착되는 메모리 용량이 증가함에 따라 이 문제는 점점 개선될 수 있음.
- 정보보호 기능 없음
 - 과거에는 R을 백엔드 서버로 사용하여 계산을 수행하는 것도 불가능했으나, 아마존 웹 서비스 클라우드 플랫폼에서 가상 컨테이너를 사용하는 등의 기술이 개발되면서 보안 문제는 개선되었음.
- 프로그램 자체의 한국어 기능을 제공하지 않음.

GUI의 종류

- R GUI
- R Studio
- Microsoft Visual Studio
- R Commander
- 그외의 GUI 및 IDE(통합개발환경)

R GUI

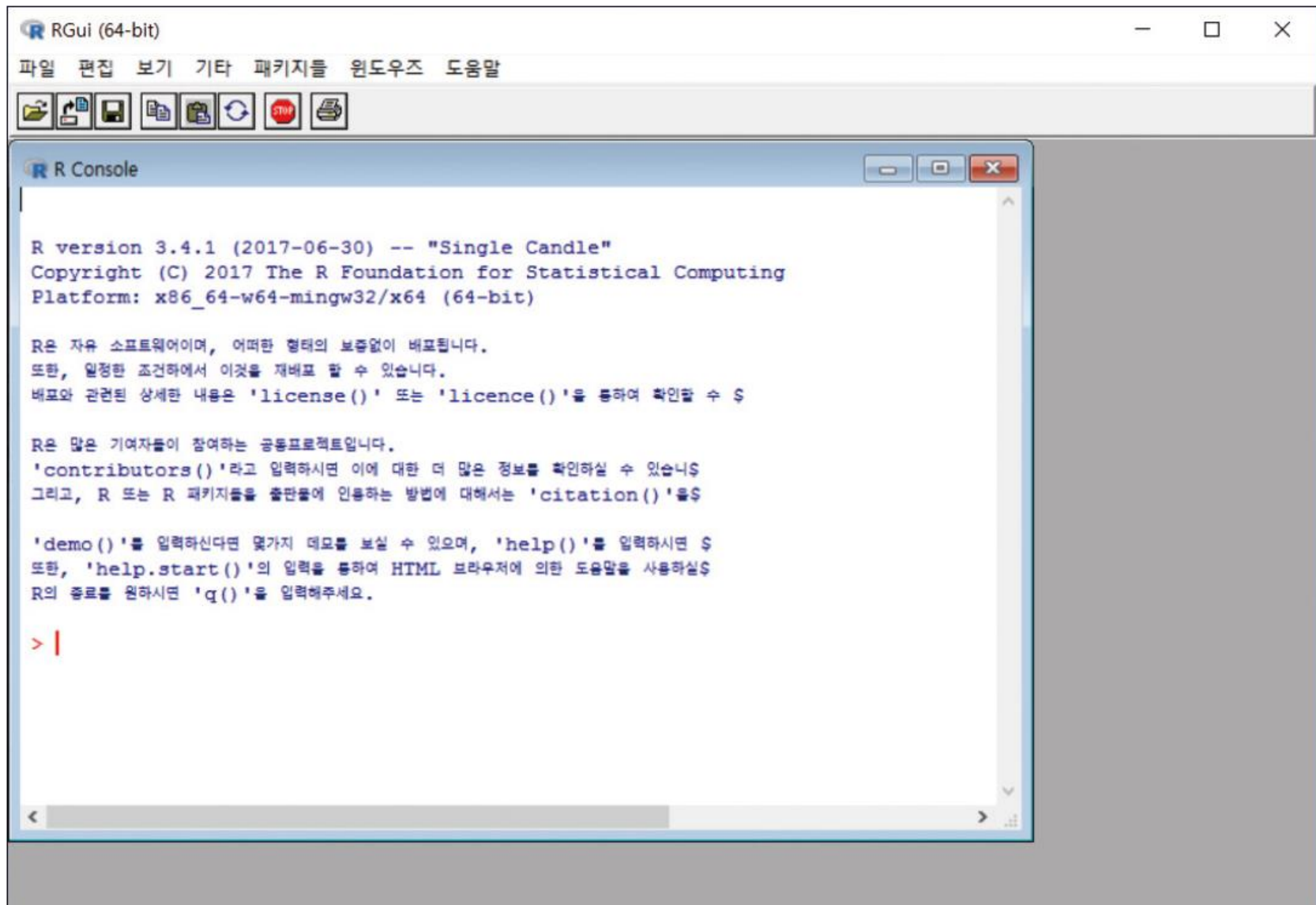


그림 2.1 윈도우에서의 표준 R 인터페이스

R Studio

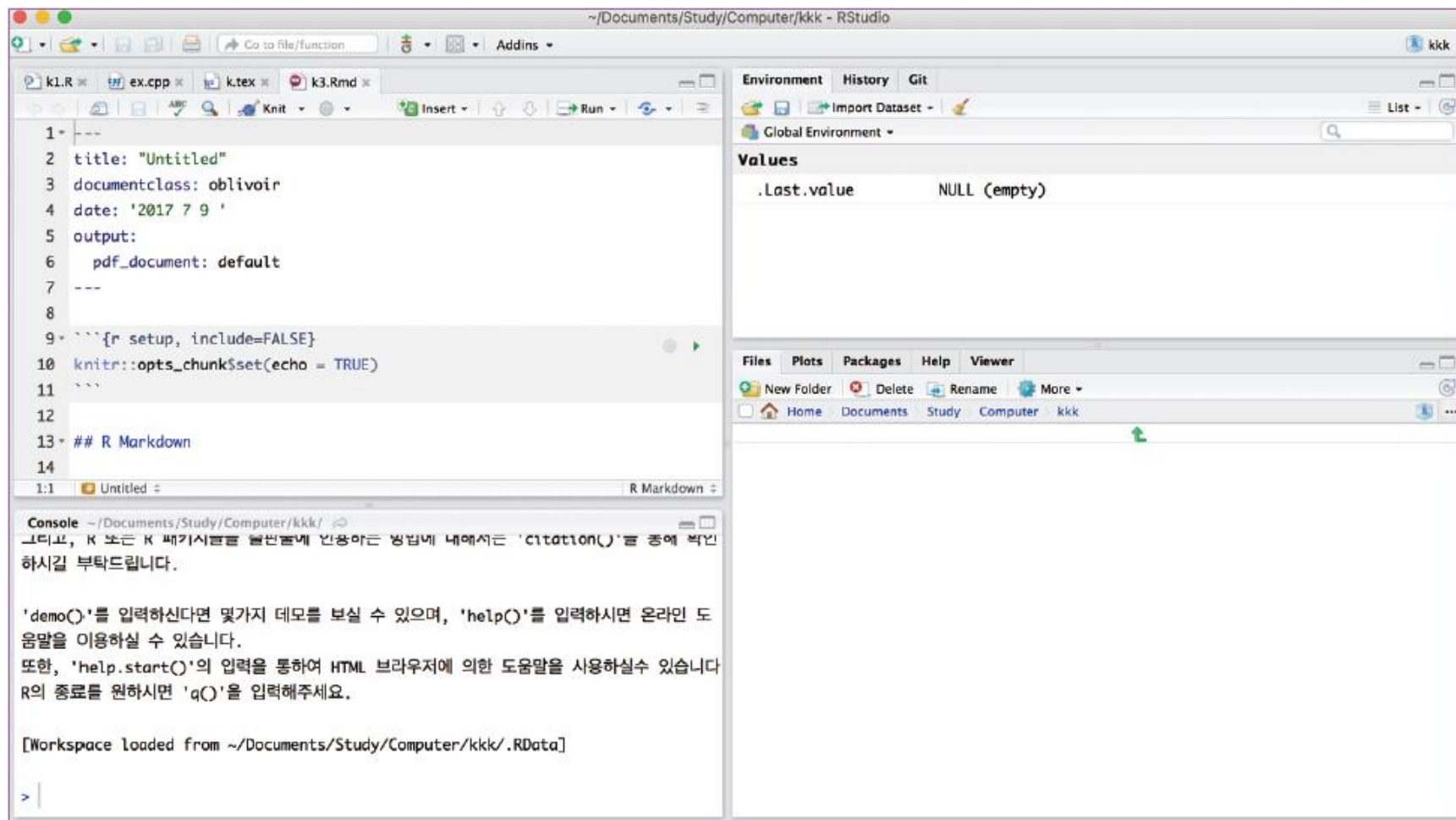


그림 2.3 RStudio의 일반적인 레이아웃

Microsoft Visual Studio

Microsoft Visual Studio interface showing R code, Variable Explorer, R Data table, and R Plot.

Code Editor (ggplot2 - example.R):

```
10 mtcars$am <- factor(mtcars$am, levels = c(0, 1),
11   labels = c("Automatic", "Manual"))
12 mtcars$cyl <- factor(mtcars$cyl, levels = c(4, 6, 8),
13   labels = c("4cyl", "6cyl", "8cyl"))
14
15 # Kernel density plots for mpg
16 # grouped by number of gears (indicated by color)
17 qplot(mpg, data = mtcars, geom = "density", fill = gear, alpha = I(.5),
18   main = "Distribution of Gas Mileage", xlab = "Miles Per Gallon",
19   ylab = "Density")
20
21 s <- sample()
22 # Scatterplot sample(x, size, replace = FALSE, prob = NULL)
23 # in each fa sample takes a sample of the specified size from the elements of x using either with
24 # facets = or without replacement.
25 # xlab = "M x: Either a vector of one or more elements from which to choose, or a positive integer.
26 See Details.
```

Variable Explorer:

Name	Value	Class	Type
mtcars	32 obs. of 11 variables	data.frame	list
@.Data	List of 11	list	list
@names	chr [1:11] "mpg" "cyl" "disp"	character	character
@row.names	chr [1:32] "Mazda RX4" "Ma	character	character
[1]	"Mazda RX4"	character	character
[2]	"Mazda RX4 Wag"	character	character
[3]	"Datsun 710"	character	character
[4]	"Hornet 4 Drive"	character	character
[5]	"Hornet Sportabout"	character	character
[6]	"Valiant"	character	character
[7]	"Duster 360"	character	character
[8]	"Merc 240D"	character	character

R Data: base::GlobalEnv\$mtcars:

	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21.0	6cyl	160.0	110	3.90	2.620
Mazda RX4 Wag	21.0	6cyl	160.0	110	3.90	2.875
Datsun 710	22.8	4cyl	108.0	93	3.85	2.320
Hornet 4 Drive	21.4	6cyl	258.0	110	3.08	3.215
Hornet Sportabout	18.7	8cyl	360.0	175	3.15	3.440
Valiant	18.1	6cyl	225.0	105	2.76	3.460
Duster 360	14.3	8cyl	360.0	245	3.21	3.570
Merc 240D	24.4	4cyl	146.7	62	3.69	3.190
Merc 230	22.8	4cyl	140.8	95	3.92	3.150
Merc 280	19.2	6cyl	167.6	123	3.92	3.440
Merc 280C	17.8	6cyl	167.6	123	3.92	3.440
Merc 450SE	16.4	8cyl	275.8	180	3.07	4.070

R Plot:

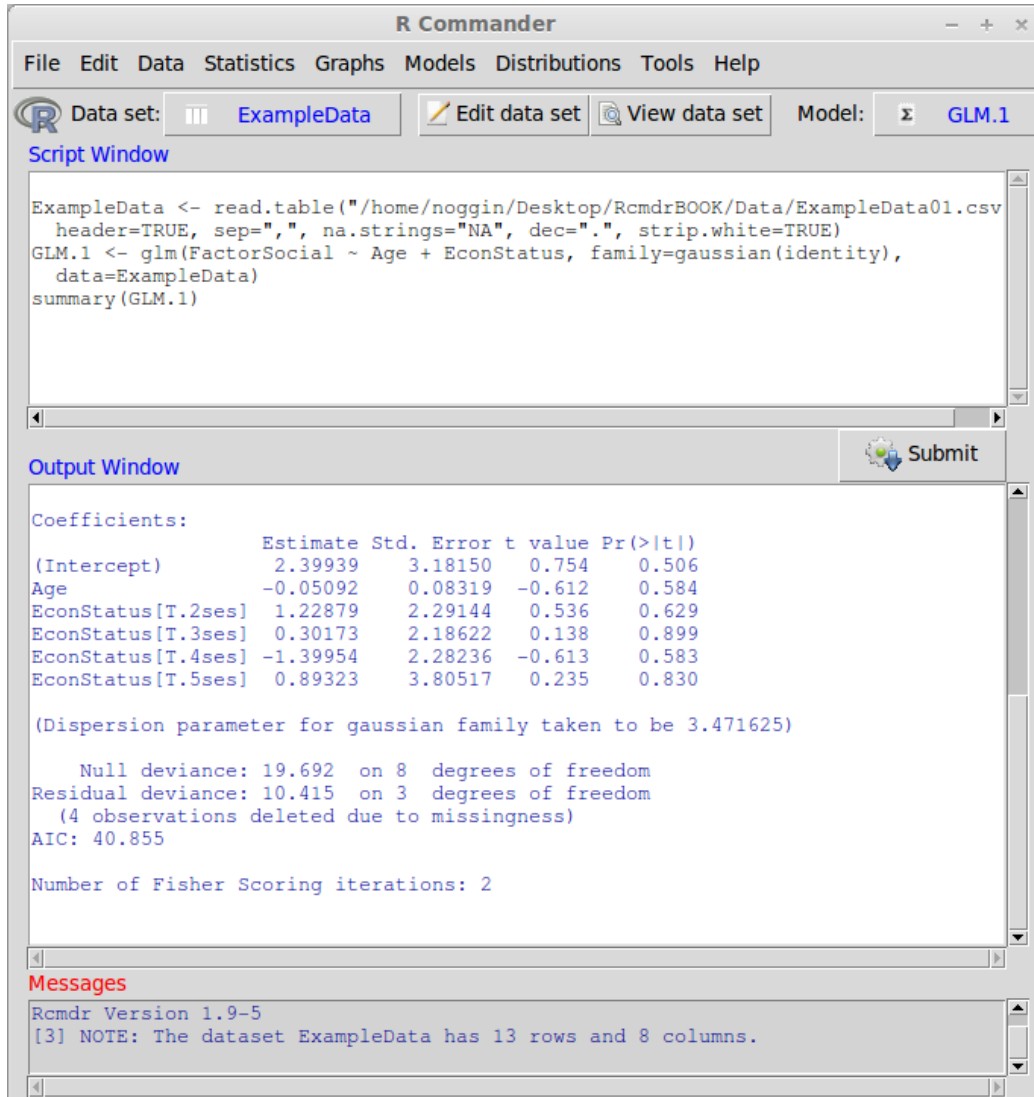
Distribution of Gas Mileage

gear

- 3gear
- 4gear
- 5gear

R Commander

- <https://www.rcommander.com/>



The screenshot shows the R Commander application window. The title bar is "R Commander". The menu bar includes File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, and Help. Below the menu bar, there are buttons for "Data set: ExampleData", "Edit data set", "View data set", and "Model: GLM.1".

The **Script Window** contains the following R code:

```
ExampleData <- read.table("/home/noggin/Desktop/RcmdrBOOK/Data/ExampleData01.csv",
  header=TRUE, sep=";", na.strings="NA", dec=".", strip.white=TRUE)
GLM.1 <- glm(FactorSocial ~ Age + EconStatus, family=gaussian(identity),
  data=ExampleData)
summary(GLM.1)
```

The **Output Window** displays the results of the GLM model. It includes a table of coefficients, the dispersion parameter, deviance statistics, and the number of Fisher Scoring iterations.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.39939	3.18150	0.754	0.506
Age	-0.05092	0.08319	-0.612	0.584
EconStatus[T.2ses]	1.22879	2.29144	0.536	0.629
EconStatus[T.3ses]	0.30173	2.18622	0.138	0.899
EconStatus[T.4ses]	-1.39954	2.28236	-0.613	0.583
EconStatus[T.5ses]	0.89323	3.80517	0.235	0.830

(Dispersion parameter for gaussian family taken to be 3.471625)

Null deviance: 19.692 on 8 degrees of freedom
Residual deviance: 10.415 on 3 degrees of freedom
(4 observations deleted due to missingness)
AIC: 40.855

Number of Fisher Scoring iterations: 2

The **Messages** window shows the following messages:

```
Rcmdr Version 1.9-5
[3] NOTE: The dataset ExampleData has 13 rows and 8 columns.
```

그외의 GUI 및 기타

- Eclipse
- Sublime Text
- Emacs
- Vim
- LyX
- jEdit
- Kate
- ConTEXT
- TextMate
- Atom
- WinEdt
- Tinn-R
- Notepad++
- Architect

주요 기능

- 통계 분석
- 데이터 마이닝
- 빅데이터 분석
- GIS
- 웹 크롤링(Web Crawling)
- 텍스트 마이닝(Text Mining)
 - 워드 클라우드(word cloud)
 - 감성분석
- 소셜 네트워크 분석(SNA: Social Network Analysis)
- 기계학습
- Reproducible Research
- Shiny를 이용한 웹 애플리케이션 개발