The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.1.087
33(1), 87–105

Applications of R package for statistical engineering

Dae-Heung Jang[a,1]

[a]Department of Statistics, Pukyong National University

(Received December 27, 2019; Revised February 5, 2020; Accepted February 5, 2020)

Abstract

Statistical engineering contains the design of experiments, quality control/management, and reliability engineering. R is a free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. R package has many functions and libraries for statistical engineering. We can use R package as a useful tool for statistical engineering. This paper shows the applications of R package for statistical engineering and suggests a R Task View for statistical engineering.

Keywords: statistical engineering, R package, design of experiments, quality control/ management, reliability engineering

1. 서론

통계학은 데이터를 수집, 정리, 분석하고 의사결정을 하는 과학으로 정의된다. 통계공학(statistical engineering)은 크게 3가지 파트인 실험계획법, 품질관리/품질경영, 신뢰성공학으로 구성된다. 이 3가지 파트는 일반적인 통계분야와는 구분되는 독특한 영역을 이루고 있다. 그러나 넓게 보면 통계공학은 다음 Table 1.1과 같이 3개의 도메인으로 나눌 수 있다. R은 무료로 개방되어 있는 통계패키지로서 통계모형, 통계 계산 및 통계 그래픽 관련 라이브러리가 2019년 12월 26일 현재 무려 15,361개가 존재할 정도로 방대하다. 이러한 패키지들이 방대하여 R core team에서는 topic별로 41개의 CRAN Task Views를 제시하고 있다. 최근에 'Teaching Statistics'라는 Task Views가 신설되었다. 우리는 이러한 R 패키지를 통계공학을 위한 기본 통계패키지로 유용하게 사용할 수 있다. 본 논문에서는 기본 도메인과 중요 도메인을 중심으로 R 패키지들 사이의 상호 의존구조를 살펴봄으로써 어떠한 패키지들이 중요한 역할을 하며 중요 R 함수들은 어떠한 함수들이 있는 지를 정리해보고자 한다. 본 논문에서 2절에서는 통계공학의 기본 도메인별로 R 패키지의 응용에 대하여 살펴보고 3절에서는 통계공학의 주요 도메인별로 R 패키지의 응용에 대하여 살펴본다. 4절에서는 결론을 맺는다.

2. 통계공학 기본 도메인에서의 R 패키지 응용

PageRank 알고리즘은 구글 검색에서 사용하는 방법으로서 검색 엔진 결과 웹사이트의 순위 (중요도)를 매기는 알고리즘이다. 이 알고리즘을 사용하여 R 패키지들의 순위 (중요도) 중 1위에서 50위까지 막대그래프로 그려보면 Figure 2.1과 같다. 통계공학 기본 도메인에서 중요한 위치를 차지하는 R 패키지들이 상당히 많이 포함되어 있음을 알 수 있다.

This work was supported by a Research Grant of Pukyong National University (2019).
[1]Department of Statistics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, Korea. E-mail: dhjang@pknu.ac.kr

Table 1.1. Domains and fields in statistical engineering

Domain	Field
Basic	Data wrangling
	Exploratory data analysis
	Statistical inference
Main	Design of experiments
	Quality control/ quality management
	Reliability engineering/ survival analysis
Advanced	Regression analysis
	Time series
	Statistical computing
	Machine learning

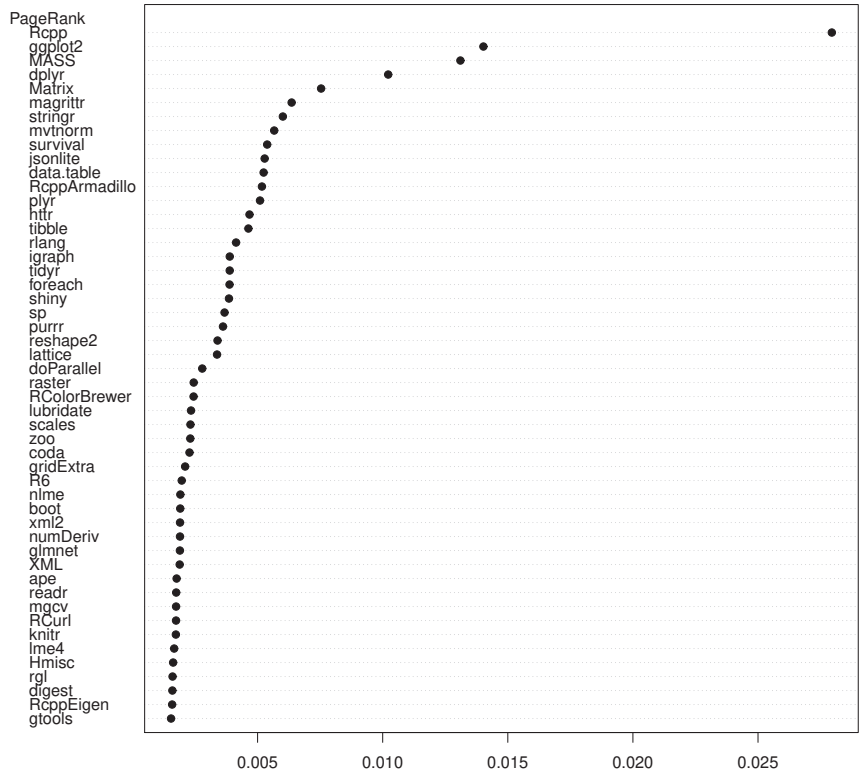

Figure 2.1. Ranking for R packages.

R 패키지 사이에는 상호 의존구조를 갖는다. R 홈페이지에 가서 특정 R 패키지를 선택하면 이 패키지 전반에 대한 소개가 있고 이 패키지와 관련된 패키지들 사이의 종속성에 대한 언급이 있다. 'Imports'는 가장 중요한 종속성으로서 언급된 패키지는 설치하는 데 필요한 재귀종속성(recursive dependencies)을 가진 완전한 패키지이다. 특정 패키지가 'Imports'에 속한 패키지들이 없으면 작동을 하지 않음을 의미한다. 'Suggests'에 언급된 패키지는 특정 패키지의 종속성을 몇 가지 기능으로 제한시키는 패키지이고 'Enhances'에 언급된 패키지는 특정 패키지의 성능을 개선시키는 패키지이

Table 2.1. The meaning of R package dependency

Dependency	Meaning
Depends	**'Depends'** is used to indicate dependency on a particular version of R, and on packages that are to be loaded (with library()) whenever your package is loaded. If you expect that users would want to load that other package whenever they loaded yours, then you should include the package name here. But this is now relatively rare.
Imports	**'Imports'** is used for packages that are needed by your package but that don't need to be loaded with library().
Suggests	**'Suggests'** is for packages that aren't really necessary, but that you're using in your examples, vignettes, or tests. Any package listed in **'Imports'** will need to be installed with your package, while packages listed in **'Suggests'** do not need to be installed with your package.
Enhances	**'Enhances'** is for packages that could improve the performance but not strictly required.

다. 'Suggests'나 'Enhances'에 속한 패키지들은 있으면 특정 패키지에서 활용하기 때문에 좋지만, 특정 패키지의 주요 기능을 수행하는 데에는 이상이 없는 패키지들이라는 것을 의미한다. 'Depends'는 R의 특정 버전을 요구한다. Table 2.1은 R 라이브러리 사이의 종속성에 대한 의미를 나타내는 표이다(http://kbroman.org/pkg_primer/pages/depends.html).

R의 miniCRAN 패키지의 pkgDep() 함수를 이용하면 관심있는 R 패키지간의 종속 구조를 알 수 있다 (Na, 2017). 이러한 R 패키지간의 종속 구조를 시각화함으로써 우리는 R 패키지간의 종속 구조를 한눈에 파악할 수가 있고 R 패키지들 사이의 상호 의존구조를 살펴봄으로써 어떠한 패키지들이 중요한 역할을 하는지를 알 수 있다.

2.1. 데이터 다루기를 위한 R 패키지의 응용

Boehmke (2016)는 데이터 다루기 (data wrangling)를 다음과 같이 정의한다.

'It's the ability to take a messy, unrefined source of data and wrangle it into something useful. It's the art of using computer programming to extract raw data and creating clear and actionable bits of information for your analysis. Data wrangling is the entire front end of the analytic process and requires numerous tasks that can be categorized within the get, clean, and transform components.'

Table 2.2는 데이터 다루기의 내용을 나타내고 있다 (Boehmke, 2016).

데이터 다루기는 데이터사이언스에서의 데이터 측면을 다루기 위한 10 가지 단계로 구성된다 (Williams, 2017).

1. Data ingestion, 2. Data review, 3. Data cleaning, 4. Variable Roles, 5. Feature selection, 6. Missing data, 7. Feature creation, 8. Preparing the Metadata, 9. Preparing for model building, 10. Save the dataset

기존의 통계공학에서는 데이터 다루기에 대하여 큰 관심을 두지 않았으나 4차 산업혁명과 인공지능 시대에 빅데이터를 다루기 위해서는 데이터 다루기에 중점을 두어야 한다. 데이터 다루기 관련 중요 R 패키지로서는 dplyr, magrittr, stringr, tibble, reshape2가 있다. Table 2.3은 데이터 다루기, 파이프 연산자, 테이블들의 결합, 문자열 연산에 대한 R 함수와 R 패키지를 보여준다.

Table 2.2. Contents for data wrangling

Process	Contents
Data Cleaning	Data Cleaning before and after Collecting Data
Working with Different Types of Data	Dealing with Numbers, Character Strings, Regular Expressions, and Dates
Managing Data Structures	Managing Vectors, Lists, Matrices, Data Frames, and Missing Values
Importing, Scraping, and Exporting Data	Importing Data, Scraping Data, Exporting Data
Creating Efficient and Readable Code	Functions, Loop Control Statements, and Pipe Operators
Shaping and Transforming Data	Reshaping Data with library(tidyr) Transforming Data with library(dplyr)

Table 2.3. R functions and R packages for data wrangling, pipe operators, combining multiple tables, and string operation

Subject	R functions	R packages
data wrangling	str(), glimpse(), select(), filter(), mutate(), arrange(), summarize() with group_by(), merge(), subset(), aggregate(), stack(), unstack(), transpose(), reshape(), sort()	library(dplyr), library(FSelector), library(lubridate), library(magrittr), library(mdsr), library(mosaic), library(plyr), library(readr), library(scales), library(stringi), library(stringr), library(tibble), library(tidyr), library(tidyverse), library(xtable)
pipe operators	%>%, %<>%, %T>%	
combining multiple tables	inner_join(), left_join(), right_join(), full_join(), semi_join(), anti_join()	
'apply' family	apply(), lapply(), sapply(), mapply(), tapply()	
string operation	grep(), nchar(), paste(), substr(), strsplit(), gregexpr(), gsub()	

dplyr 패키지는 데이터 다루기를 위한 R 패키지로서 매우 중요한 라이브러리이다. dplyr 패키지는 파이프 연산자(%<>%, %$%, %>%, %T%)와 결합하여 각 종 데이터 처리(split-apply-combine(분할-적용-병합))을 다루는 함수들(select(), filter(), mutate(), group_by(), summarize() 등)로 구성되어 있다.

어떤 패키지의 영향력은 reverse dependencies (reverse depends, reverse imports, reverse suggests, reverse enhances)로 확인할 수 있는 데 특히 'reverse imports'가 매우 중요하다. library(dplyr)는 'reverse imports' 패키지가 무려 1,409개, 'reverse suggests' 패키지가 368개, 'reverse depends' 패키지가 77개, 'reverse enhances' 패키지가 3개이다. 수많은 데이터 어낼리틱스 관련 패키지들을 구동하기 위해서는 dplyr 패키지가 필수적인 패키지인 것을 확인할 수 있다. Figure 2.2는 dplyr 패키지와 다른 패키지와의 종속성을 나타내는 다이아그램이다. 파이프 연산자와 관련이 있는 magrittr 패키지가 dplyr 패키지와 연관이 있음을 알 수 있다. 또한 tidy data 관련 패키지인 tibble 패키지도 dplyr 패키지와 연관이 있음을 알 수 있다.

Figure 2.3은 데이터 다루기를 위한 패키지들의 종속성을 나타내는 다이아그램이다. dplyr 패키지와 ggplot2 패키지가 양대 산맥을 형성하고 있음을 알 수 있다. 데이터 다루기는 데이터 시각화와 매우 밀접한 연관성이 있음을 시사한다.

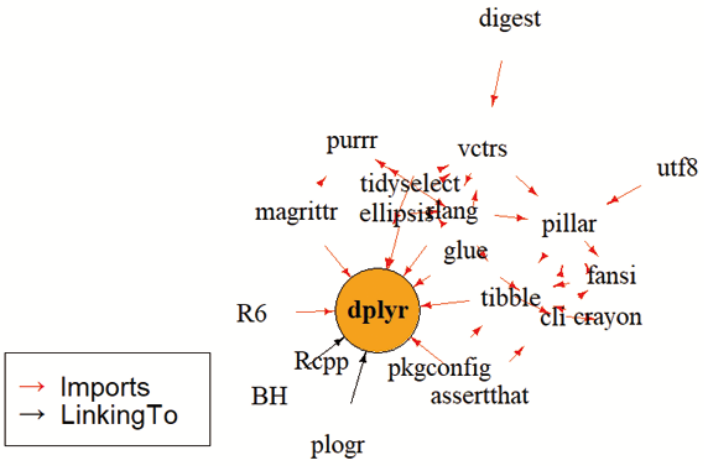

Figure 2.2. Diagram which display R package dependency linked to 'dplyr' package.

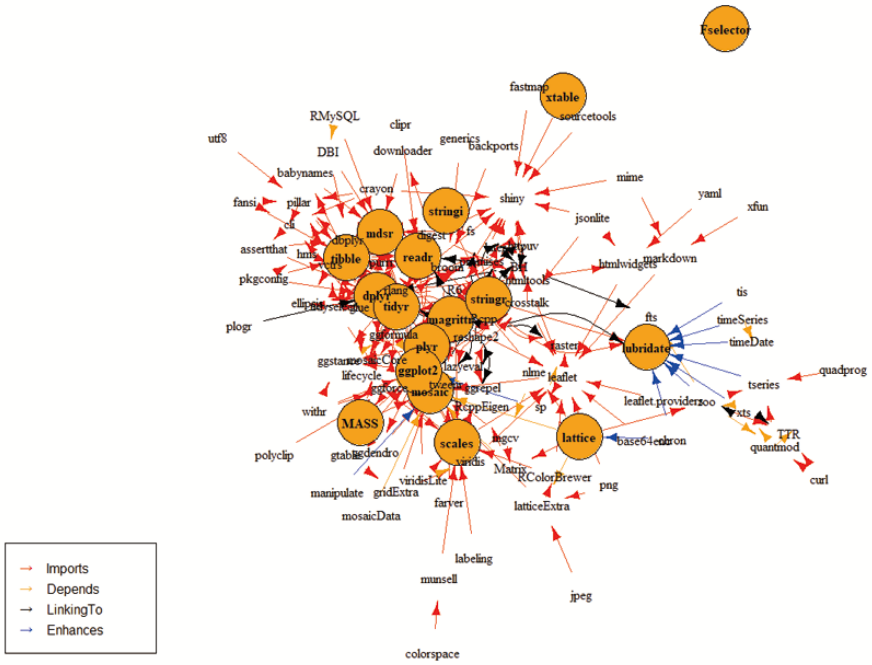

Figure 2.3. R package dependency graph for data wrangling.

2.2. 탐색적자료분석을 위한 R 패키지의 응용

Table 2.4는 탐색적자료분석을 위한 R 함수와 R 패키지를 나타낸다. 탐색적자료분석에서는 데이터 시각화가 매우 중요하다. 데이터 시각화 방법 중 동적 그래픽스는 여러 가지 동적 테크닉(brushing, identification, scaling, subset selection, line segments, interactive drawing, dragging points, rotation, animation, animation with control bars)을 포함한다. 웹 상의 그래픽스 구현을 위해서는 shiny 패키

Table 2.4. R functions and R packages for exploratory data analysis

Subject	R functions	R packages
Comparison of several groups - numerical measures	max(), min(), quantile(), mean(), median(), var(), sd(), IQR(), mad(), range(), sum(), prod(), cumsum(), cumprod(), cummax(), cummin(), colMeans(), colSums(), rowMeans(), rowSums(), summary(), table()	library(animation), library(beanplot), library(diagram), library(ggplot2), library(hdrcde), library(hexbin),
Comparison of several groups - graphical meathods	stripchart(), stem(), boxplot(), hist(), vioplot(), hdr.boxplot(), bpplot(), beanplot(), densityplot()	library(Hmisc), library(iplots), library(lattice),
Re-expression	exp(), log(), log10(), sqrt(), boxcox(), transform()	library(latticeExtra), library(MASS),
Probability plots	qqnorm(), qqline(), qqmath(), qqplot(), fitdistr()	library(misc3d), library(plot3D),
Two-way data table & two-way frequency table	medpolish(), barplot(), dotchart(), mosaicplot (), assocplot(), fourfoldplot()	library(plotly), library(plotrix),
Exploration of time series	smooth(), decompose(), ecf(), ccf()	library(RColorBrewer),
Robust regression	lm(), rlm(), lqs()	library(rggobi),
Exploration of bivariate data	cov(), cor(), plot(), jitter(), symbols(), stars(), face(), density(), kde2d(), hexbin(), curve(), matplot(), sunflowerplot(), contour(), image(), filled.contour(), hdr.boxplot.2d()	library(rgl), library(scatterplot3d), library(shiny), library(tabplot) library(vcd)
Exploration of multivariate data	persp(), plot3d(), persp3d(), surface3d(), cylinder3d(), scatteplot3d(), persp3D(), scatter3D(), hist3D(), surf3D(), coplot(), xyplot(), pairs(), tableplot()	library(vcdExtra), library(vioplot),
Dynamic graphics	plot3d(), iplot(), ggobi()	

지나 plotly 패키지가 필요하고 2차원 동적 그래픽스를 위해서는 rggobi 패키지나 iplots 패키지가 필요하다. 애니메이션을 위해서는 animation 패키지가 필요하고 3차원 동적 그래픽스를 그리기 위해서는 rgl 패키지가 필요하다. CRAN Task View 'Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization'에서는 core package로서 ggplot2, lattice, plotrix, RColorBrewer, rgl, vcd를 제시하고 있다. 통계공학 중요 도메인에서 언급할 품질관리(quality control; QC) 중 QC 활동 전개에 필요한 기본적인 QC 수법으로서 우리는 주로 다음과 같은 7가지 도구(seven tools of QC)를 사용한다.

1. 히스토그램, 2. 특성요인도(causes-and-effects diagram), 3. 파레토그림(Pareto diagram), 4. 체크시이트(check sheet), 5. 산점도(scatter disgram), 6. 층별(stratification), 7. 각 종 그래프(graphs)

7가지 도구들 모두 데이터 시각화와 밀접한 관계를 갖는다. QC 신 7가지 도구(new seven tools of QC)는 기존의 7가지 도구를 보완하기 위하여 제안되어 산업현장에서 사용되고 있다.

1. 친화도, 2. 연관도, 3. 계통도, 4. 매트릭스도, 5. 매트릭스 데이터 해석, 6. Process decision program chart (PDPC), 7. 애로우 다이아그램

5번의 매트릭스 데이터 해석은 다변량통계에서 다루는 주성분그림(principal component plot)을 의미한다. 신 7가지 도구들 또한 모두 데이터 시각화와 밀접한 관계를 갖는다.

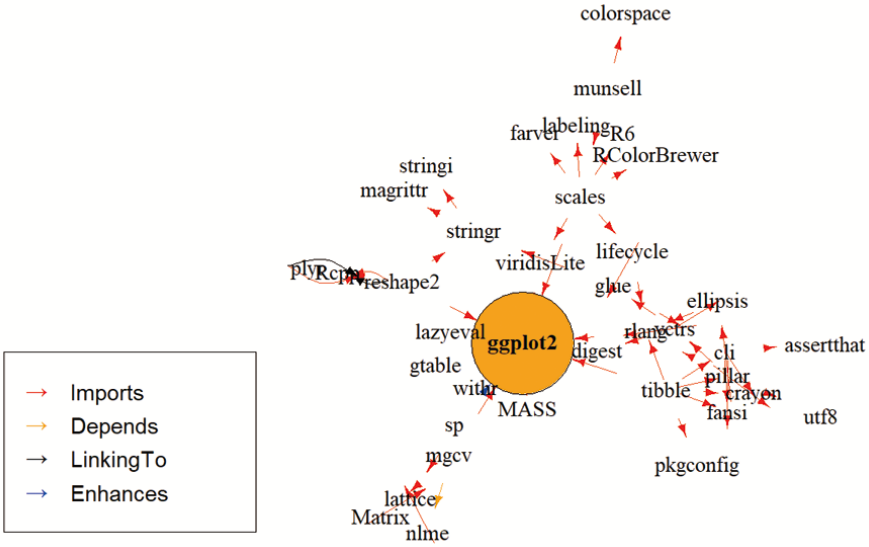

Figure 2.4. Diagram which display R package dependency linked to 'ggplot2' package.

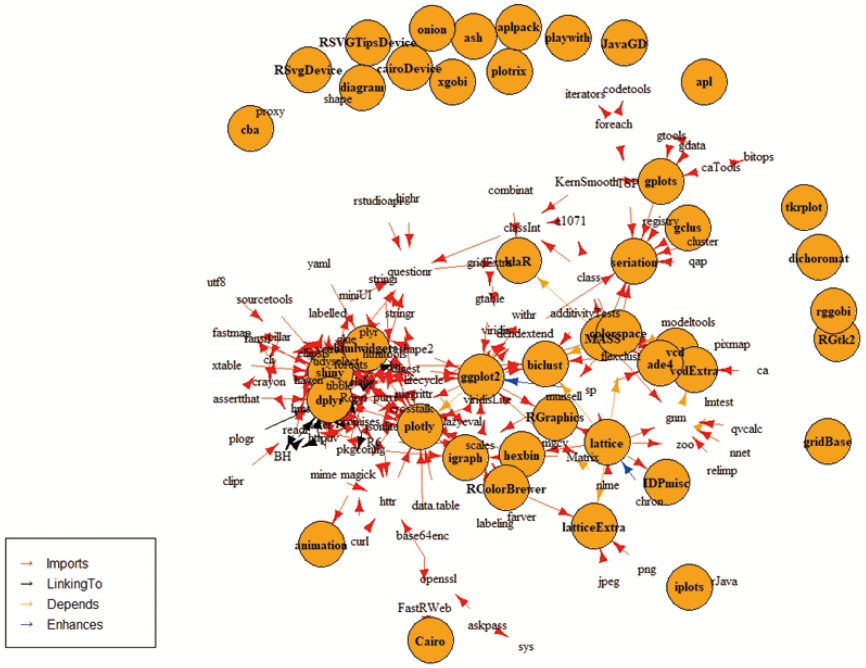

Figure 2.5. R package dependency graph for data visualization.

ggplot2 패키지는 통계그래픽스를 위한 R 패키지로서 매우 중요한 라이브러리이다. Wilkinson (2005)의 'Grammar of Graphics'의 철학을 바탕으로 Hadley Wickham이 만든 그래픽스 라이브러리로서 객체지향 디자인과 증분형식을 사용한다. 총 12개의 함수군(Plot creation, Geoms, Statistics, Scales, Co-

ordinate systems, Faceting, Position adjustments, Annotation, Fortify, Themes, Aesthetics)으로 구성되어 있다. 우리는 ggplot2 패키지를 통하여 다양한 기능을 조합하여 복잡하고 미려한 플롯을 그릴 수 있다. ggplot2 패키지는 'reverse imports' 패키지가 무려 1,544개, 'reverse suggests' 패키지가 728개, 'reverse depends' 패키지가 318개, 'reverse enhances' 패키지가 1개이다. 수많은 통계그래픽스 및 데이터 어낼리틱스 관련 패키지들을 구동하기 위해서는 ggplot2 패키지가 필수적인 패키지인 것을 확인할 수 있다. ggplot2 패키지의 그래픽스 철학을 따라 만들어진 통계그래픽스 패키지들(패키지 명칭이 gg 로 시작하는 패키지들)이 107개가 있다. Figure 2.4는 ggplot2 패키지와 다른 라이브러리와의 종속성을 나타내는 다이아그램이다. 통계그래픽스를 위한 또 다른 라이브러리인 lattice 패키지가 ggolot2 패키지와 연관이 있음을 알 수 있다. 또한 데이터 변환 관련 패키지인 reshape2 패키지와 stringr 패키지가 ggplot2 패키지와 연관이 있음을 알 수 있다. 통계학 책 (Modern Applied Statistics with S(4th edition, 2002)) 관련 패키지인 MASS 패키지가 ggplot2 패키지와 매우 연관이 있는 것도 흥미로운 사실이다.

Figure 2.5는 데이터 시각화를 위한 패키지들의 종속성을 나타내는 다이아그램이다. ggplot2 패키지와 lattice 패키지가 양대 산맥을 형성하고 있음을 알 수 있다. 웹 상의 플롯과 관련이 깊은 plotly 패키지와 shiny 패키지도 이 두 패키지들과 관계가 있음을 알 수 있다.

예제 2.1: 다음은 qcc 패키지에 있는 피스톤 링 자료 (부분군[로트]의 갯수가 40개이고 부분군의 크기는 5)를 이용하여 기본 패키지와 ggplot2 패키지를 이용하여 상자그림을 각각 그리는 R script이고 Figure 2.6은 gplot2 패키지를 이용하여 그린 상자그림이다. 각 로트에 대응되는 상자 안의 별표는 산술평균을 나타낸다.

```
install.packages("qcc"); install.packages("ggplot2")
library(qcc); library(ggplot2)
data(pistonrings); attach(pistonrings)
str(pistonrings)
summary(pistonrings)
# box plot with mean
boxplot(diameter ~ sample,xlim=c(1,40))
points(tapply(diameter,sample,mean),pch=10,col="red")
# box plot using 'ggplot2' package
ggplot(data=pistonrings, aes(x=factor(sample), y=diameter)) +
geom_boxplot(fill="yellow") +
stat_summary(fun.y="mean", geom="point", shape=8, size=3) +
theme(axis.title.x = element_text(family ='sans' , face = 2, size=20)) +
theme(axis.title.y = element_text(family ='sans' , face = 2, size=20))
```

3. 통계공학 주요 도메인에서의 R 패키지 응용

3.1. 실험계획법을 위한 R 패키지의 응용

실험계획법은 통계공학에서 매우 중요한 자리를 차지하는 연구 영역이다. 실험계획법은 완전확률화계획법, 요인실험, 확률화블럭계획법, 분산성분, 부분요인실험, 불완비/교락블럭계획법, 분할법, cross-over designs, 반복측정계획, 직교계획, definitive screening designs, covering array (CA), computer-

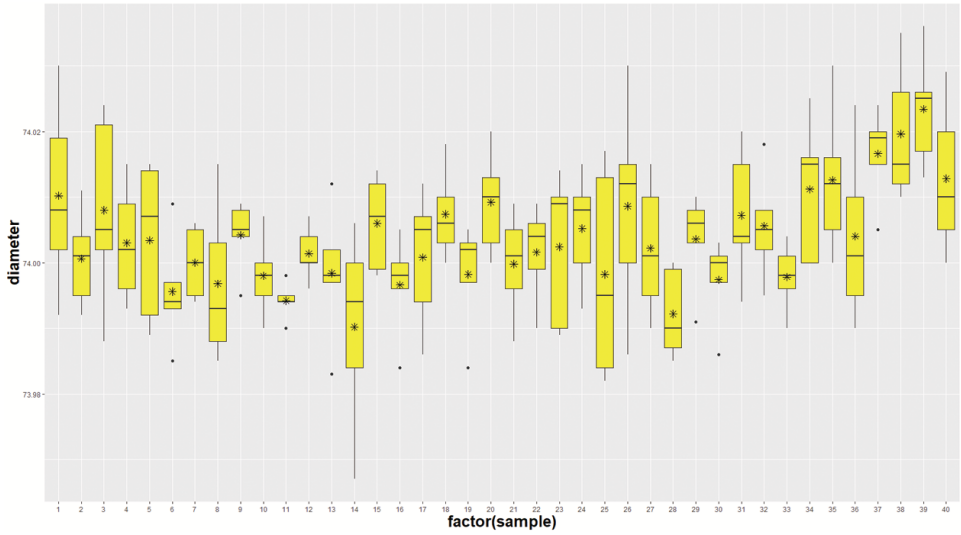

Figure 2.6. Box plots for piston rings data.

generated designs, 초포화계획, 반응표면분석, 혼합물실험계획법, robust parameter designs 등으로 구성된다. 실험계획법은 내용상 크게 실험계획 파트와 통계분석 파트 두 가지 영역이 있다. 통계분석 파트에서 회귀분석과 분산분석(선형회귀, 일반화선형모형, 비선형회귀, 혼합모형, 축소추정량, 회귀진단 포함)과 연결이 된다. CRAN Task View 'Design of Experiments (DoE) & Analysis of Experimental Data'에서는 core package로서 agricolae, AigDesign, conf.design, Crossdes, DoE.base, DoE.wrapper, FrF2, rsm, skpr을 제시하고 있다. Table 3.1은 실험계획법을 위한 R 함수와 R 패키지를 제시한다.

FrF2 패키지는 부분요인실험법을 위한 R 패키지로서 실험계획법에서 DoE.base 패키지와 더불어 매우 중요한 패키지이다. Figure 3.1은 FrF2 패키지와 다른 패키지와의 종속성을 나타내는 다이아그램이다. DoE.base 패키지가 FrF2 패키지와 연관이 있음을 알 수 있다. 흥미로운 사실은 사회연결망분석 도구인 igraph 패키지가 FrF2 패키지와 연관이 있다는 사실이다.

Figure 3.2는 실험계획법을 위한 패키지들의 종속성을 나타내는 다이아그램이다. FrF2 패키지와 DoE.base 패키지를 중심으로 하나의 그룹을 형성하고 이 그룹은 데이터 시각화에서 중요한 역할을 하는 ggplot2 패키지와 데이터 변환에서 중요한 역할을 하는 dplyr 패키지와 함께 세 개의 축을 이룸을 알 수 있다. 실험계획법을 위한 분석을 위해서는 데이터 시각화와 데이터 변환이 필수적으로 필요함을 알 수 있다. 몇 가지 흥미로운 사실이 있다. 첫째, 이러한 그룹들 사이에 MASS 패키지가 위치해 있다. 둘째, survival 패키지(생존분석 관련 패키지), quantreg 패키지(분위수회귀(quantile regression) 관련 패키지), lme4 패키지(선형혼합모형 관련 패키지) 같은, 생존분석 모형이나 회귀분석 모형 관련 패키지들과 연결되어 있다. 셋째, 기계학습모형의 하나인 신경망모형 패키지와도 연결되어 있음을 알 수 있다. nnet 패키지는 신경망 모형 중 은닉층이 하나인 shallow learning 모형인 feed-forward neural networks이다.

예제 3.1: Jones과 Nachtsheim (2011, 2013)은 Definitive Screening Design (DSD)라는 새로운 종류의 선별 계획을 제안하였다. 다음 daewr 패키지를 사용하는 R script를 통하여 우리는 3개의 수준을 갖

Table 3.1. R functions and R packages for experimental designs

Subject	R functions	R packages
Completely randomized designs	factor(), lm(), summary(), summary.lm(), aov(), anova(), fit.contrast(), contr.poly(), boxcox(), polr(), Fpower1(), TukeyHSD(), SNK.test(), glht()	library(AlgDesign), library(agricolae), library(BsMD),
Factorial designs	expand.grid(), factor(), data.frame(), lm(), coef(), lsmeans(), model.tables(), estimable(), contr.poly(), predict(), aggregate(), subset(), glm(), interaction.plot(), Fpower1(), Fpower2(), Tukey1df(), fullnormal(), LGB(), LenthPlot(), BsProb(), Gaptest(), contourPlot(), effectPlot(), facDesign(), interactionPlot(), normalPlot(), pbDesign(), wirePlot()	library(car), library(conf.design), library(crossdes), library(daewr), library(desirability), library(DoE.base), library(DoE.wrapper),
Randomized block design	factor(), sample(), levels(), design.rcbd(), data.frame(), aov(), summary(), contr.poly(), summary.aov(), do.call(), apply(), tapply(), lm(), Fpower(), model.tables(), TukeyHSD(),design.lsd()	library(FrF2), library(GAD), library(gmodels), library(lsmeans),
Variance components	data.frame(), aov(), summary(), model.matrix(), as.matrix(), lmer(), profile(), qchisq(), qf(), pf(), vci(), estimable(), lsmeans(), anova(), qgamma(), qnorm(), ranef()	library(leaps), library(lme4), library(MASS), library(multcomp),
Fractional factorial designs	data.frame(), lm(), summary(), FrF2(), aliases(), add.response(), LGB(), generators(), fold.design(), gen.factorial(), optFederov(), pb(), castfr(), regsubsets(), summary(), OptPB(), Altscreen(), ModelRobust(), step(), instep(), bstep(), oa.design(), anova(), DefScreen(), alliasTable(), fracDesign()	library(mixexp), library(pbkrtest), library(qualityTools), library(rsm), library(skpr), library(Vdgraph)
Incomplete and confounded block design	BIBsize(), optBlock(), aov(), lsmeans(), summary(), design.cyclic(), FrF2(), add.response(), lm(), coef(), halfnorm(), FrF2(), gen.factorial(), fac.design(), mod(), as.factor(), oa.design(), lm(), Anova(), lsmeans()	
Split-plot designs	expand.grid(), optBlock(), FrF2(), as.fixed(), as.random(), aov(), gad(), lmer(), anova(), factor(), lsmeans(), coef(), summary(), fullnormal(), aliases(), lm()	
Crossover and repeated measure designs	lm(), Anova(), lsmeans(), qt(), data.frame(), lmer(), summary(), williams(), paste(), tapply(), as.fixed(), as.random(), anova(), factor(), aov(), model.tables()	
Definite screening designs	designs DefScree(), colormap()	
Response surface designs	ccd(), Vdgraph(), ccd.pick(), bbd(), transform(), optFederov(), Compare2FDS(), data.frame(), rsm(), summary(), nls(), persp(), contour(), steepest(), constrOptim(), dMax(), dMin(), dTarget(), dOverall(), gen.factorial(), optBlock(), class(), lmer(), EEw2s3(), rsmDesign(), starDesign()	
Mixture experiments	SLD(), DesignPoints(), lm(), summary(), MixModel(), MixturePlot(), EffPlot(), Xvert(), optFederov(), optBlock(), crvtave(), data.frame(), Fillv(), subset(), expand.grid(), merge(), constrOptim(), lmer(), transform(), contourPlot3(), mixDesign()	

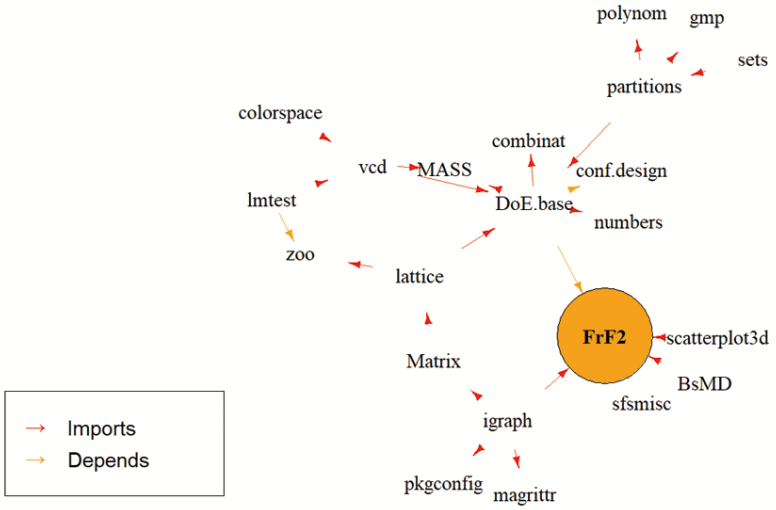

Figure 3.1. Diagram which display R package dependency linked to 'FrF2'package.

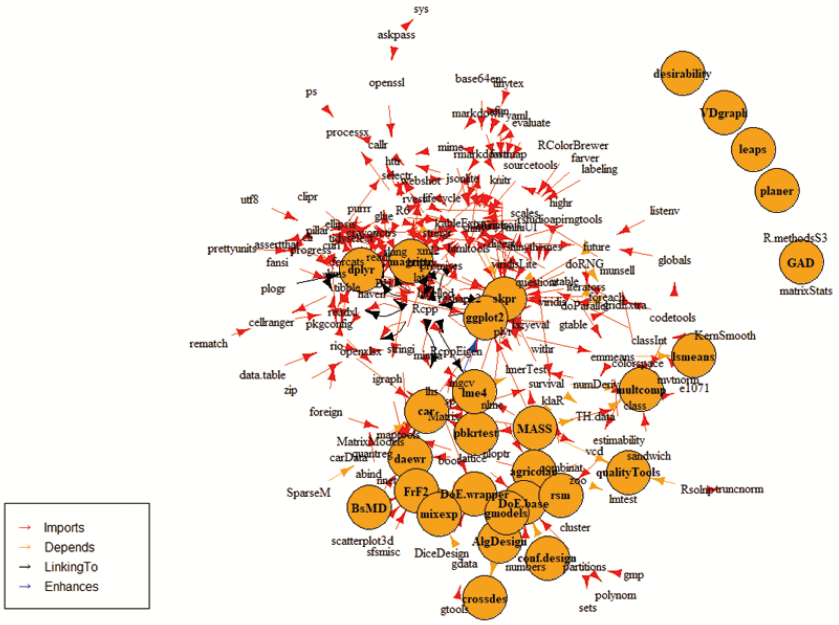

Figure 3.2. R package dependency graph for experimental designs.

는 8개의 양적 인자이면서 17개의 실험계획점을 갖는 DSD를 만들고 이 DSD의 2차 모형행렬을 대상으로 상관계수를 구한 후 Figure 3.3처럼 color map을 그릴 수 있다. 이 color map을 통하여 주효과는 2인자 순수효과 및 2인자 혼합효과에 대하여 교락되어 있지 않고 2인자 순수효과들끼리의 교락 정도는 약함을 알 수 있다. 또한 특정 2인자 효과들끼리는 교락되지 않는 특이한 패턴을 형성함을 알 수 있다.

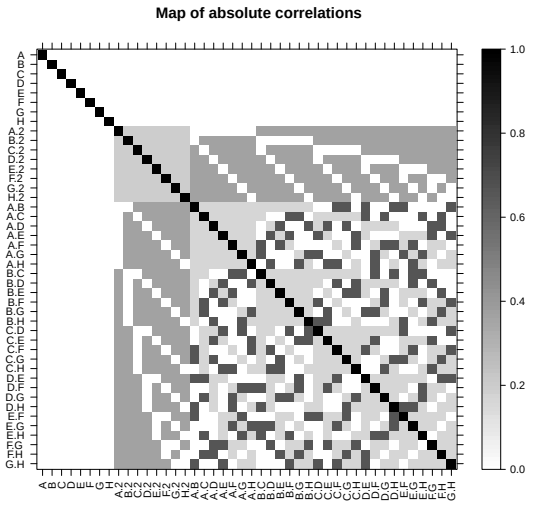

Figure 3.3. Color map of 17-run DSD for 8 quantitative factors.

```
install.packages("daewr"); library(daewr)
defscr <- DefScreen(m=8)
y<-runif(nrow(defscr),0,1)
test<-model.matrix(lm(y~(.)^2,data=defscr))
q <- I(test[,2:9])^2
colnames(q) <- c("A^2", "B^2", "C^2", "D^2", "E^2", "F^2", "G^2", "H^2")
defs <- data.frame(cbind(test[, 2:9],q,test[,10:37 ]))
colormap(defs,mod=1)
```

다음 예제는 혼합물 실험계획법(experiments of mixtures) 관련 예제이다. 혼합물은 단체(simplex)를 형성하게 되고 이러한 구조는 compositional data에서도 나타난다. 이러한 혼합물 예제를 R 패키지를 이용하여 해석할 수 있다.

예제 3.2: Juan 등 (2006)은 3개의 구성성분을 갖는 혼합물 실험 ('polvdat'데이터)을 제시하였다. 이 논문에서는 반응변수로서 polvoron 과자에 대한 소비자 만족도(consumer acceptance)를 고려하였다. polvoron 과자는 3가지 구성성분인 설탕($x1$), 땅콩($x2$), 버터($x3$)로 만들어지고 다음과 같은 제약조건을 갖는다.

$$x1 + x2 + x3 = 1$$
$$0.00 <= x1 <= 0.80$$
$$0.10 <= x2 <= 0.95$$
$$0.05 <= x3 <= 0.50$$

다음은 이러한 제약조건 하에서 daewr 패키지를 이용하여 특수 3차 회귀모형 예측값을 구하기 위한 R script와 시행 결과를 나타낸다.

```
install.packages("daewr"); library(daewr)
data(polvdat)
sqm = lm(y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3 - 1, data = polvdat)
summary(sqm)
MixturePlot(des = polvdat, mod = 4, lims = c(0, .8, .1, .95, .05, .50), constrts = T,
pseudo = T)
```

```
> sqm = lm(y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3 -1, data = polvdat)
> summary(sqm)
Call:
lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3 -
1, data = polvdat)
Residuals:
      1       2       3       4       5       6       7       8
-0.17957 -0.02142 0.03359 -0.12009 0.14423 -0.20166 0.09631 -0.14312
      9      10      11      12
0.20803 0.01123 0.29123 -0.11877
Coefficients:
       Estimate   Std.Error  t value   Pr(>|t|)
x1       4.4259     0.4483    9.873   0.000182 ***
x2       3.5181     0.3079   11.427   8.99e-05 ***
x3       1.2367     1.6150    0.766   0.478400
x1:x2    6.9004     2.0179    3.420   0.018846 *
x1:x3    8.9528     4.1427    2.161   0.083071 .
x2:x3    5.3135     3.4988    1.519   0.189310
x1:x2:x3 25.5460   11.2023    2.280   0.071499 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2374 on 5 degrees of freedom
Multiple R-squared: 0.9992, Adjusted R-squared: 0.9981
F-statistic: 920.9 on 7 and 5 DF, p-value: 1.815e-07
```

Figure 3.4는 이러한 제약조건 하에서 특수 3차 회귀모형 예측값을 나타내는 등고선도를 나타낸다.

3.2. 품질관리/ 품질경영을 위한 R 패키지의 응용

품질관리와 품질경영에 대하여 위키피디아(https://en.wikipedia.org/wiki/Quality_control, https://en.wikipedia.org/wiki/Quality_management)는 다음과 같이 정의하고 있다.

Quality control is a process by which entities review the quality of all factors involved in production. Quality management ensures that an organization, product or service is consistent. It has four main components: quality planning, quality assurance, quality control and quality improvement.

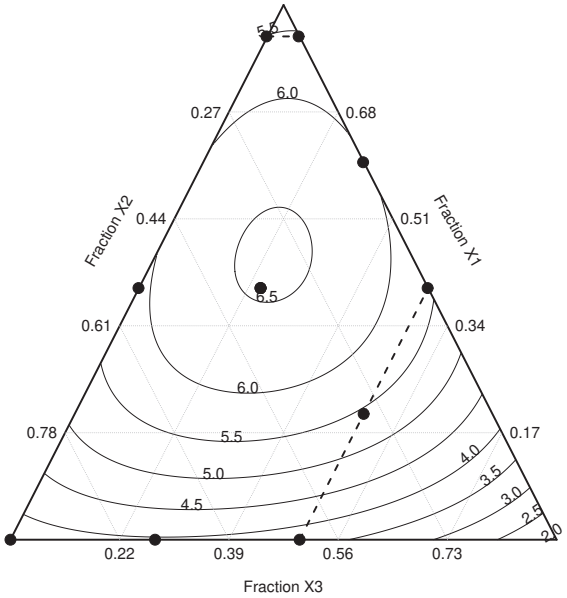

Figure 3.4. Contour plot in restricted pseudo component region under the special cubic model

Table 3.2. R functions and R packages for quality control/ quality management

Subject	R functions	R packages
Probability distributions	dhyper(), dbinorm(), dmultinom(), dpois(), dnbinom(), dgeom(), dunif(), dnorm(), dexp(), dt(), dchisq(), dbeta(), dgamma(), dweibull(), dcauchy(), dlnorm(), ppPlot(), qqPlot()	library(AcceptanceSampling), library(cpm), library(Dodge), library(ggQC), library(IQCC),
Cause-and-effect diagrams & Pareto charts	cause.and.effect(), ss.ceDiag(), paretoChart(), pareto.chart(),	library(mnspc), library(msqc),
Acceptance sampling	find.plan(), acc.samp(), OC2c(), assess(), oc.curves()	library(mpcv), library(MSQC),
Control charts	qcc(), ewma(), cusum(), mqcc(), ss.cc(), qic()	library(plan), library(plotrix),
Capability analysis	cp(), process.capability(), ss.ca.cp(), ss.ca.cpk(), ss.ca.z()	library(qAnalyst), library(qcc),
Gage R & R	gageRR(), averagePlot(), compPlot(), errorPlot(), gageLin(), gageRRDesign(), whiskerPlot()	library(qcr), library(qicharts), library(qualityTools),
Nonlinear profiles	climProfiles(), plotProfiles(), outProfiles(), plotControlProfiles()	library(SixSigma), library(spc),
Graphical method	mvPlot(), snPlot(), dotPlot()	library(spcadjust), library(tolerance)

품질관리 및 품질경영 중 통계공학과 관련된 내용은 대개 계량치 및 계수치의 추검정, 상관분석, 계수형 및 계량형 샘플링 검사, 계수형 및 계량형 관리도, 공정관리, Gage R & R 등으로 구성된다. 품질관리 및 품질경영 관련 중요 패키지로서는 qcc, qualityTools, SixSigma가 있다. ggQC 패키지는 ggplot2 패

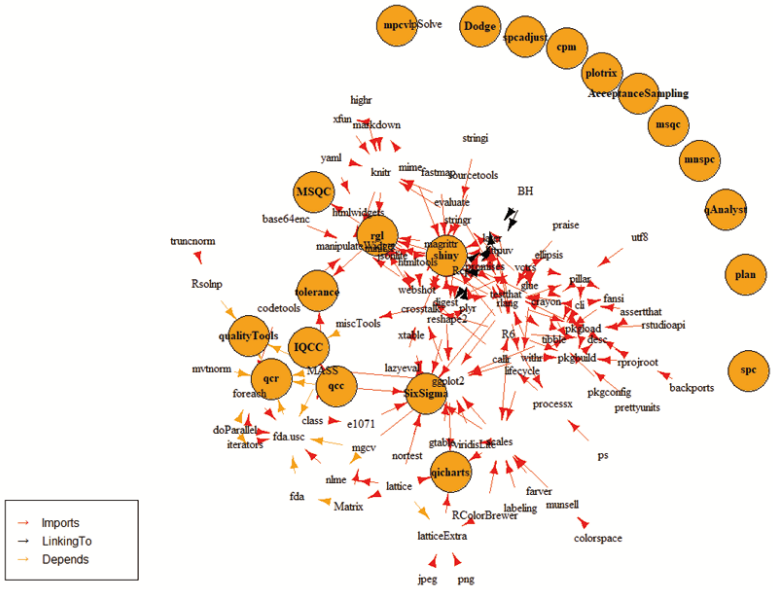

Figure 3.5. R package dependency graph for quality control/ quality management.

키지의 그래픽스 철학 하에 만든 관리도 패키지이다. Table 3.2는 품질관리와 품질경영을 위한 R 함수와 R 패키지를 제시한다.

Figure 3.5는 품질관리 및 품질경영을 위한 패키지들의 종속성을 나타내는 다이아그램이다. six sigma 관련 패키지인 SixSigma가 3차원 데이터 시각화에서 중요한 역할을 하는 rgl 패키지와 웹상에서의 데이터 시각화 구현 관련 shiny 패키지와 함께 세 개의 축을 이룸을 알 수 있다. 식스 시그마를 위한 분석을 위해서는 데이터 시각화가 필수적으로 필요함을 알 수 있다. 또한 품질관리 및 품질경영을 위한 패키지들은 기계학습 관련 패키지들인 e1071 패키지 (naive Bayes, SVM, fuzzy clustering 같은 기계학습 모형 관련 패키지), rpart 패키지 (의사결정나무 모형 관련 패키지) 등과 연결되어 있고 회귀모형 관련 패키지들인 nlme 패키지 (선형, 비선형혼합모형 관련 패키지), mgcv 패키지 (GAM모형 관련 패키지), functional data analysis 관련 패키지인 fda 패키지와도 연결되어 있다.

예제 3.3: (**예제 3.1에서 계속**) 다음은 패키지 qcc, SixSigma, qualityTools를 사용하여 예제 1의 피스톤 링 자료에 대한 계량형 관리도를 그리고 공정능력지수를 산출해 보는 R script이다.

```
install.packages("qcc"); install.packages("SixSigma"); install.packages("qualityTools")
library(qcc); library(SixSigma); library(qualityTools)
data(pistonrings); str(pistonrings)
attach(pistonrings); summary(pistonrings)
GROUP <- qcc.groups(data = pistonrings$diameter, sample = pistonrings$sample)
# X-bar chart
qcc(GROUP, type = "xbar")
# Range chart
qcc(GROUP, type = "R")
# Process capability analysis (library(qcc))
```

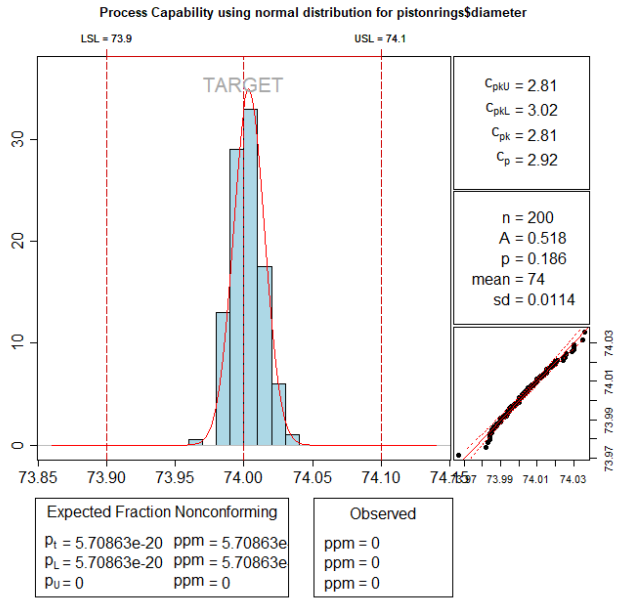

Figure 3.6. Process capability analysis for piston rings data(using 'qualityTools' package).

```
QCC <- qcc(GROUP, type = "xbar", plot = FALSE)
process.capability(QCC, spec.limits = c(73.9, 74.1), target = 74)
# Process capability analysis (library(SixSigma))
ss.study.ca(pistonrings$diameter, LSL = 73.9, USL = 74.1, Target = 74)
# Process capability analysis (library(qualityTools))
cp(x = pistonrings$diameter, lsl = 73.9, usl = 74.1, target = 74)
```

Figure 3.6은 qualityTools 패키지를 사용하여 피스톤 링 자료에 대한 공정능력분석을 실시한 결과를 나타낸다.

3.3. 신뢰성공학을 위한 R 패키지의 응용

신뢰성공학은 무생물인 재료, 기구, 제품, 소프트웨어의 수명에 주관심이 있고 생존분석은 환자의 생존시간에 주관심이 있다. 신뢰성공학과는 달리 생존분석에서는 절단자료 (censored data)가 자주 발생하므로 비모수추론을 주로 행한다. 또한 신뢰성공학에서는 통계모형으로서 와이블분포를 중심으로 전개되나 생존분석은 감마분포를 중심으로 전개된다. 그럼에도 불구하고 신뢰성공학은 생존분석과 아주 유사한 통계분석을 행한다. 품질경영에서 healthcare 영역이 최근의 최대 관심사인 바 신뢰성공학과 생존분석의 연구는 서로 상생하는 융합 학문 분야가 형성되고 있다. 생존분석 모형 관련 패키지는 Hmisc 패키지를 통해 신경망모형 관련 패키지인 nnet 패키지와도 연결되어 있다. CRAN Task View 'Survival Analysis'에서는 core package로서 cmprsk, eha, mstate, muhaz, rms, survival, timereg를 제시하고 있다. Table 3.3은 신뢰성공학 및 생존분석을 위한 R 함수와 R 패키지를 제시한다.

Figure 3.7은 신뢰성공학 및 생존분석을 위한 패키지들의 종속성을 나타내는 다이아그램이다. 생존분석

Table 3.3. R functions and R packages for reliability engineering/ survival analysis

Subject	R functions	R packages
Reliability analysis	wblr(), contour.wblr(), wblr.conf(), wblr.fit(), plot.wblr(), plot_contour(), AbPval(), Fmbound(), getCCC2(), getPPP(), hrbu(), rba(), LLln(), LLw(), LRbounds(), lslr(), MLEcontour(), mlefit(), MLEln2p(), MLEln3p(), MLEw2p(), MLEw3p(), MRRln2p(), MRRln3p(), MRRw2p(), MRRw3p(), weibayes(), weibayes.mle(), wp.test()	library(aftgee), library(ciTools), library(cmprsk), library(dhglm), library(eha), library(frailtyHL), library(GFGM.copula),
Software reliability	duane(), littlewood.veall(), moranda.geometric(), musa.kumamoto(), mvf.duane(), mvf.mor(), mvf.musa(), mvf.ver.lin(), mvf.ver.quad(), duane.plot(), littlewood.veall.plot(), moranda.geometric.plot(), musa.kumamoto.plot(), rel.plot(), total.plot()	library(hglm), library(HGLMMM), library(joint.Cox), library(mstate), library(muhaz), library(Reliability),
Survival analysis	Surv(), survreg(), survfit(), survdiff(), summary.survfit(), coxph(), cox.zph(), summary.coxph(), survfit.coxph(), calculate_ranks(), confint_betabinom(), confint_fisher(), delta_method(), john_method(), kaplan_method(), nelson_method(), mixmod_regression(), ml_estimation(), mr_method(), rank_regression(), plot_conf(), plot_mod(), plot_mod_mix(), plot_pop(), plot_prob(), plot_prob_mix(), predict_prob(), predict_quantile(), dist[mcs]_delay_register(), dist[mcs]_delay_report(), dist[mcs]_milege(), Kaplan.meier.location(), Lifedata.MLE(), lifetime.mle(), psev(), summary.lifedata.MLE(), survest.cph(), survest.psm(), survfit.cph(), residual.cph(), survplot(), hglm(), dhglmfit(), HGLMfit(), frailtyHL()	library(rms), library(rstpm2), library(simexaft), library(SPREDA), library(survival), library(timereg), library(weibullness), library(WeibullR), library(weibulltools)
accelerated failure time model	add_ci(), add_pi(), aft(), aftgee(), simexaft()	

관련 survival 패키지, 데이터 변환에서 중요한 역할을 하는 dplyr 패키지, 그리고 웹상에서의 데이터 시각화 구현 관련 shiny 패키지 및 plotly 패키지가 세 개의 축을 이룸을 알 수 있다. 신뢰성공학 및 생존분석을 위해서는 데이터 변환 및 데이터 시각화가 필수적으로 필요함을 알 수 있다.

4. 결론

우리는 R 패키지들을 통계공학을 위한 기본 통계패키지로 유용하게 사용할 수 있다. 본 논문에서 통계공학의 기본 및 주요 도메인에 관련된 R 함수와 R 패키지들을 대상으로 R 패키지간의 종속 구조를 시각화함으로써 우리는 R 패키지간의 종속 구조를 한 눈에 파악할 수가 있었다. 이를 통하여 어떠한 패키지들이 중요한 역할을 하며 R 함수들은 어떠한 함수들이 있는 지를 정리해보았다. R 홈페이지에는 통계공학 관련 CRAN Task Views가 아직 존재하지 않으므로 통계공학에서 R 패키지를 사용하는 데 도

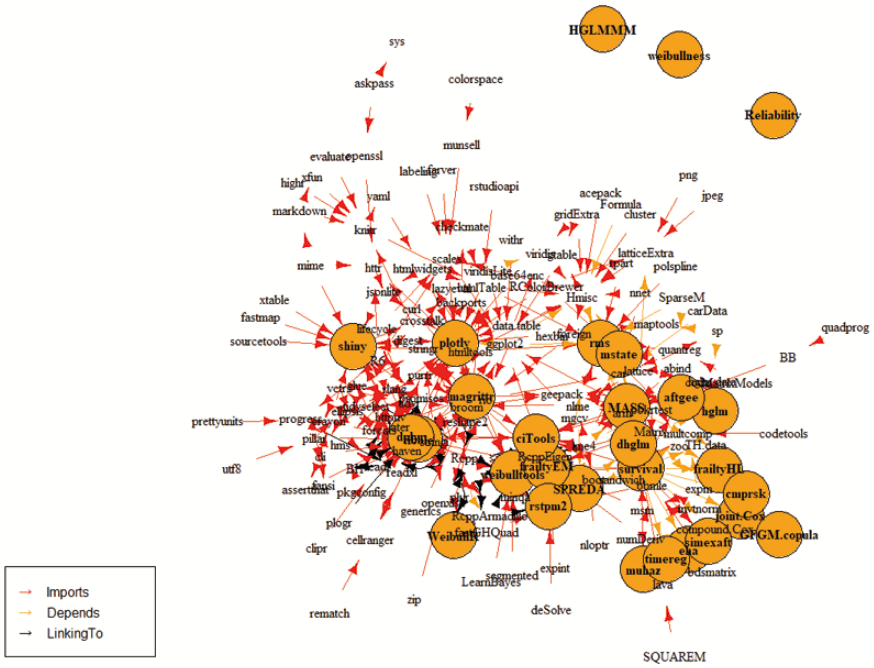

Figure 3.7. R package dependency graph for reliability engineering/ survival analysis.

움을 주기 위한 가이드로서 통계공학 관련 CRAN Task Views를 만들자는 제안을 한다.

References

Baumer, B. S., Kaplan, D. T., and Horton, N. J. (2017). *Modern Data Science with R*, CRC Press, New York.

Boehmke, B. C. (2016). *Data Wrangling*, Springer, New York.

Cano, E. L., Moguerza, J. M., and Corcoba, M. P. (2015). *Quality Control with R*, Springer, New York.

Cano, E. L., Moguerza, J. M., and Redchuk, A. (2012). *2. Six Sigma with R*, Springer, New York.

Jones, B. and Nachtsheim, C. J. (2011). A class of three-level designs for definitive screening in the presence of second-order effects, *Journal of Quality Technology*, **43**, 1–15.

Jones, B. and Nachtsheim, C. J. (2013). Definitive screening designs with added two-level categorical factors, *Journal of Quality Technology*, **45**, 121–129.

Juan, E. M. S., Edra, E. V., Sales, J. M., Lustre, A. O., and Resurreccion, A. V. A. (2006). Utilization of peanut fines in the optimization of peanut polvoron using mixture response methodology, *International Journal of Food Science and Technology*, **41**, 768–774.

Lawson, J. (2015). *Design and Analysis of Experiments with R*, CRC Press, New York.

Montgomery, D. C. (2013a). *Design and Analysis of Experiments* (8th ed), John Wiley, Singapore.

Montgomery, D. C. (2013b). *Introduction to Statistical Quality Control* (7th ed), John Wiley, New York.

Na, J. H. (2017). *R Data Mining*, Free Academy, Seoul.

Qiu, P. (2014). *Introduction to Statistical Process Control*, CRC Press, New York.

Wilkinson, L. (2005). *The Grammar of Graphics*, Springer, New York.

Williams, G. J. (2017). *The Essentials of Data Science*, CRC Press, New York.

통계공학을 위한 R 패키지 응용

장대흥[a, 1]

[a]부경대학교 통계학과

(2019년 12월 27일 접수, 2020년 2월 5일 수정, 2020년 2월 5일 채택)

요 약

통계공학은 실험계획법, 품질관리/품질경영, 신뢰성공학으로 구성된다. R은 무료로 개방되어 있는 통계패키지로서 통계모형, 통계 계산 및 통계 그래픽 관련 패키지가 방대하다. 우리는 이러한 R 패키지를 통계공학을 위한 기본 통계패키지로 유용하게 사용할 수 있다. 본 논문에서는 통계공학을 위한 R 패키지 응용을 살펴보고 통계공학 관련 CRAN Task Views가 필요함을 제안하였다.

주요용어: 통계공학, R 패키지, 실험계획법, 품질관리/품질경영, 신뢰도공학

이 논문은 부경대학교 자율창의학술연구비(2019년)에 의하여 연구되었음.

[1](48513) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: dhjang@pknu.ac.kr