응용통계연구 제 20권 3호, 2007년, pp. 619–634

R에 의한 통계그래픽스 : 강의 내용 및 방법의 논의

박동련[1)]

요 약

자료분석과정에서 그래프의 이용은 필수적이라고 하겠다. 다양하게 개발된 수많은 그래픽 기법들을 적절하게 사용할 수 있다면 한 단계 업그레이드된 통계분석이 가능할 것이며, 이런 면에서 볼 때 통계그래픽스는 통계학을 전공하는 학생들에게 꼭 필요한 강좌라고 할 수 있다. 다양하게 개발된 그래픽 기법의 막강한 파워를 제대로 느끼기 위해서는 적절한 통계 소프트웨어의 선택이 매우 중요한 문제라고 할 수 있는데, 뛰어난 그래픽 기능이 있는 R을 사용하는 것이 효율적으로 다양한 그래픽 기법을 구현할 수 있는 가장 바람직한 선택이라고 하겠다. 이 논문에서는 통계 그래픽스를 R을 이용하여 구현하는 강좌를 개설하고자 하는 경우에 사용할 수 있는 적절한 교과내용을 제안하고, 어떤 방식으로 강의하는 것이 가장 효과적인지에 대한 고민을 함께 해 볼 수 있는 기회를 제공하고자 한다.

주요용어: 자료분석도구, 통계그래픽스, R.

1. 서론

자료분석과정에서 그래프의 이용은 필수적이라고 하겠다. 잘 그려진 그래프를 이용한다면 우리는 다른 어떤 분석을 이용하는 것보다도 더 명쾌하게 자료의 구조를 꿰뚫어 볼 수 있게 되는데, 이것은 곧 그래프를 이용한 자료분석방법이야말로 자료가 담고 있는 정보를 가장 잘 활용할 수 있는 분석방법임을 의미한다고 할 수 있다. 그림 1.1은 이미 많이 알려진 그래프로서 학생들에게 그래프를 이용한 자료분석이 왜 필요한지를 확실히 보여줄 수 있는 Cleveland (1993)에 의한 놀라운 발견이다.

그림 1.1에서 사용된 자료는 1930년대 초에 미네소타주의 농경학자들이 보리 종류에 따른 수확량의 차이를 비교하기 위하여 2년 동안 경작실험을 하여 얻은 자료이다. 고려된 요인으로는 6군데의 경작지와 10종류의 보리, 그리고 2년간의 경작년도이며, 반응변수는 수확량이다. 세 요인의 모든 조합에 대한 자료가 구해져 모두 $6 \times 10 \times 2 = 120$개의 관찰값이 존재한다. 이 자료는 실험계획법에 대한 Fisher의 아이디어가 적용되어 실시된 최초의 실험자료 중의 하나라는 점에서 역사적 의의가 있는 자료이며, 실제로 Fisher (1971)와 Anscombe (1981), Daniel (1976) 등의 저명한 학자들에 의하여 각각 분석된 유명한 자료이다 (Cleveland, 1993).

1) (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수
E-mail: drpark@hs.ac.kr

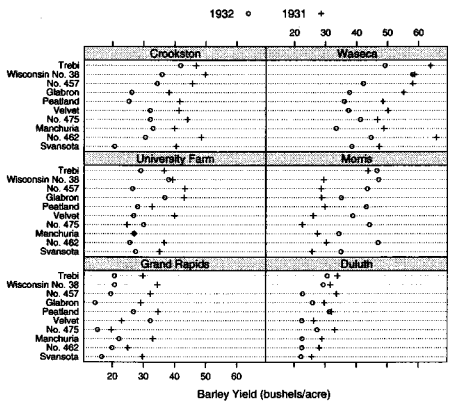

그림 1.1: 1930년대에 실시된 보리수확 실험결과 자료에 대한 다중 점그림표

이렇게 많은 학자들이 분석하였으나 이들 모두 자료에 있는 문제를 파악하지 못하였다. 문제는 Morris라는 지역의 자료에 있다. 다른 모든 경작지에서는 1931년도 수확량이 1932년도 수확량보다 많은 것으로 나타나 있는데 유독 Morris에서만 정반대의 결과가 나와있다. 6개의 경작지가 모두 같은 주에 있기 때문에 이것은 상당히 의심스러운 일이라고 할 수 있으며, Morris의 1931년과 1932년의 수확량을 바꾸면 훨씬 더 일관성이 있게 세 요인의 효과를 나타낼 수 있음이 Cleveland (1993)에 의하여 밝혀졌다.

과거의 많은 저명한 학자들이 보리 수확량 자료에 있는 문제를 미처 발견하지 못한 것은 자료의 구조를 꿰뚫어 볼 수 있는 그림 1.1과 같은 명쾌한 그래프가 그 당시에는 아직 개발되지 않았기 때문이었다. 이와 같이 그래프야말로 엄청난 위력을 발휘할 수 있는 효과적인 분석도구이며, 따라서 자료분석과정에서 적절한 그래프의 이용은 자료분석의 수준을 한단계 업그레이드 시킬 수 있음을 학생들에게 주지시켜야 할 것이다.

막강한 파워를 지닌 그래픽 기법을 효과적으로 사용하기 위해서는 적절한 소프트웨어를 선택해야 하는데, 시중에 나와있는 수많은 통계관련 소프트웨어 중에 가장 적합한 소프트웨어로 R을 들 수 있을 것이다. R이 가지고 있는 장점은 소스가 공개된 무료 소프트웨어라는 점이다. 이것은 곧 많은 사용자들이 쉽게 내용을 추가할 수 있다는 것을 의미하는 것이고, 따라서 최첨단 분석도구가 거의 실시간 수준으로 R에서 사용가능하게 되는 것이다. 또 다른 R의 장점은 뛰어난 그래픽 기능이다. R의 탁월한 그래픽 기능은 이미 여러 전문가들에 의해서 공인된 사실이며, 이러한 특징들로 인하여 통계 그래픽스는 R로 구현하는 것이 가장 효율적이라는 결론을 내릴 수 있게 된다.

이 논문에서는 통계 그래픽스를 R을 이용하여 구현하는 강좌를 개설하고자 하는 경우에 사용할 수 있는 적절한 교과내용을 제안하고, 어떤 방식으로 강의하는 것이 가장 효과적인지에 대한 고민을 함께 해 볼 수 있는 기회를 제공하고자 한다.

2. 교과내용의 제안

통계그래픽스 강좌의 적절한 강의목표는 통계자료의 분석도구로 사용할 수 있는 그래픽 기법의 습득이라고 할 수 있다. 분석도구로 그래픽 기법을 사용한다는 것은 곧 자료의 유형에 맞추어 적절한 그래픽 기법을 선택해야 한다는 것을 의미하는 것이다. 여기에서 자료의 유형은 함께 고려해야 할 변수의 개수로 구분하는 것이 그래프를 그리는 관점에서 볼 때 가장 적절하다고 하겠다. 따라서 일변량 (univariate), 이변량 (bivariate), 삼변량 (trivariate), 그리고 4변량 이상의 자료를 의미하는 초변량 (hypervariate)으로 자료의 유형을 구분하여 각 유형의 자료를 분석하는데 적합한 그래픽 기법을 차례로 소개하는 것이 바람직한 교과내용이라고 할 수 있다.

2.1. 일변량 자료를 위한 그래픽스

일변량 자료에 대하여 우리가 가장 관심을 가지고 있는 것은 자료의 분포형태일 것이다. 자료의 분포를 나타내는 방법은 범주형 자료와 연속형 자료의 경우로 구분하여야 하는데, 일변량 범주형 자료의 도수분포 형태를 나타내기 위하여 사용되는 대표적인 그래프의 목록과 대응되는 그림함수의 목록이 표 2.1에 있다. 연속형 자료의 경우에도 각 관찰값마다 라벨이 있고 그 라벨을 일종의 범주로 취급하여 그래프를 그려야 한다면 표 2.1에 있는 그래프로 속성을 나타낼 수 있다.

범주형 자료에 대한 그래프로서 가장 흔하게 접할 수 있는 것은 아마도 막대그림표와 파이그림표일 것이다. 그러나 상대적으로 잘 알려져 있지 않은 점그림표가 사실은 범주형 자료에 대한 가장 이상적인 그래프로 인정받고 있는데, 이것은 그래프에 부가적인 정보를 쉽게 추가할 수 있는 장점이 있기 때문이다. 그림 2.1은 미국 고속도로 I90이 지나는 주의 이름과 그 주에서의 I90 길이에 대한 점그림표이다. 자료 I90은 Journal of Statistical Education의 Data Archive인

```
http://www.amstat.org/publications/jse/jse_data_archive.html
```

에 있는 ushighway2.dat에서 고속도로 I90에 해당되는 13개 케이스만을 골라 구성한 것으로, 각 케이스마다 주 이름이 라벨로 붙어있는 연속형 자료이다. 따라서 범주형 자료를 위한 그래프를 이용해야 하는 경우가 된다. 추가된 세 수직선은 I90 총 길이의 5%, 10%, 그리고 15%에 해당하는 길이를 나타내기 위한 것으로 더 명확한 정보를 전하고 있다.

표 2.1: 일변량 범주형 자료에 적합한 그래프와 그림함수

그래프	그림함수
막대그림표 (Bar chart)	barplot
파이그림표 (Pie chart)	pie
점그림표 (Dot chart)	dotchart

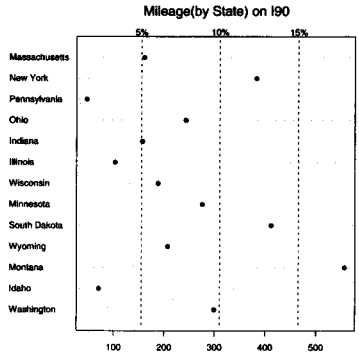

그림 2.1: I90 자료에 대한 점그림표

표 2.2: 일변량 연속형 자료에 적합한 그래프와 그림함수

그래프	그림함수
줄기-와-잎 그림 (Stem-and-leaf plot)	stem
점그림 (Dot plot)	stripchart
히스토그램 (Histogram)	hist
상자그림 (Box plot)	boxplot
커널밀도함수 추정	density

연속형 자료의 분포형태를 효과적으로 나타낼 수 있는 그래프와 대응되는 그림함수의 목록이 표 2.2에 있다. 각 그래프는 모두 나름대로의 장단점을 가지고 있는데, 매끄러운 곡선의 모습을 취하고 있는 연속형 자료의 확률밀도함수 형태를 가장 잘 나타낼 수 있는 방법은 커널밀도함수 추정량의 추정결과를 그래프로 나타내는 것이라고 하겠다. 그러나 비모수적 밀도함수 추정방법에 대해서는 학부수준에서 거의 강의가 이루어지지 않고 있는 것으로 알고 있으며, 따라서 기초적인 내용에 대한 간략한 소개는 반드시 있어야 할 것이다. 커널밀도함수 추정방법에 대한 자세한 내용은 Simonoff (1996)이나 Wand와 Jones (1995) 등에서 찾아볼 수 있다.

그림 2.2는 미국 Yellowstone 국립공원의 어떤 간헐천의 분출지속시간에 대한 밀도함수를 추정한 결과이다. 사용된 자료는 패키지 MASS에 있는 데이터 프레임 geyser의 변수인 duration이다. 히스토그램과 커널밀도함수의 추정결과를 한 그래프에 겹쳐서 나타냈다.

2.2. 이변량 자료를 위한 그래픽스

이변량 자료에 대해서는 각 변수의 분포를 효과적으로 나타내는 것뿐만이 아니라 두 변수의 분포가 같은지 혹은 다르다면 어떻게 다른지를 비교하는 것도 중요한 문제가 된다. 또

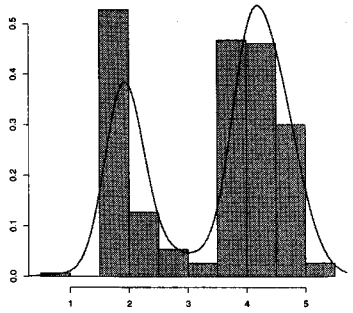

그림 2.2: 간헐천 분출지속시간에 대한 밀도함수 추정결과

표 2.3: 연속형 두 변수의 분포를 비교하기 위한 그래프와 그림함수

그래프	그림함수
나란히-서-있는 상자그림 (Side-by-side box plot)	boxplot, plot
다중 점그림 (Multiple Dot plot)	stripchart
분위수-분위수 그림 (Quantile-quantile plot)	qqplot, qqnorm

한 두 변수간의 관계를 알아보는 것도 매우 중요한 관점이 될 것이다. 우선 연속형 두 변수의 분포를 비교하기 위한 그래프와 그림함수의 목록이 표 2.3에 있다.

두 변수의 분포를 비교하는데 가장 효과적인 그래프는 분위수-분위수 그림이라고 할 수 있다. 분위수-분위수 그림을 이용하여 우리는 두 변수의 분포를 비교하는 것뿐만이 아니라 한 변수의 분포가 어떤 특정한 분포를 따르는지 여부를 확인하는 작업도 하게 되는데, 이 경우 정규분포와의 비교는 함수 qqnorm()으로 간단하게 할 수 있지만 다른 분포와의 비교를 하기 위해서는 몇가지 함수를 함께 사용해야만 할 수 있게 된다. 예를 들어 벡터 x가 자유도가 2인 카이제곱분포를 따르는지를 알아보기 위한 분위수-분위수 그림은 다음의 명령문으로 그릴 수 있다.

```
> plot(qchisq(ppoints(x),2),sort(x),xlab="",ylab="",main="Chisquare QQ plot")
```

두 변수간의 관계를 알아보기 위한 첫 번째 작업은 산점도를 그리는 것이 될 것이다. 이어서 두 변수의 관계를 가장 잘 나타내는 회귀선을 추정하여 산점도에 추가하는 것이 두 번째 작업이 되는데, 융통성이 떨어지는 모수적 회귀모형만으로는 두 변수의 관계를 나타내는 데에 종종 한계를 느낄 수 밖에 없게 된다. 따라서 자료들의 있는 그대로의 모습을 효과적으로 잘 나타내어 주는 비모수적 회귀모형을 사용하는 것이 그래픽 관점에서 본다면 훨씬 더 바람직하다고 하겠다.

데이터 프레임 faithful은 패키지 datasets에 있는 자료로서 Yellowstone 국립공원의 Old

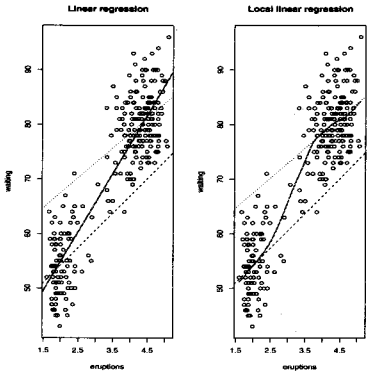

그림 2.3: 모수적 회귀모형과 비모수적 회귀모형의 적합결과

Faithful 간헐천의 분출지속시간 (eruptions)과 분출간격 (waiting)을 나타내고 있다. 자료에 대한 자세한 설명은 Azzalini와 Bowman (1990)에서 찾을 수 있다. 그림 2.3은 faithful에 있는 두 변수의 산점도와 변수 waiting을 반응변수로, eruptions을 설명변수로 하는 회귀모형을 적합시킨 결과를 그래프로 나타낸 것이다. 왼쪽 그래프의 굵은 실선은 전체 자료를 대상으로한 단순선형회귀모형의 추정결과이고, 두 점선은 변수 eruptions의 값 3.0을 경계로 하여 나누어져 있는 두 군집의 자료를 분리하여 각각 단순선형회귀모형을 적합시켜 얻은 추정결과이다. 굵은 실선의 기울기가 두 점선의 기울기와 큰 차이가 있음을 알 수 있으며, 이것은 바로 전체 자료를 대상으로 하여 추정된 회귀직선이 실제 두 변수의 관계를 정확하게 나타내지 못함을 나타내는 것이라 하겠다. 반면에 국소선형회귀모형의 추정결과를 나타내는 오른쪽 그래프의 굵은 실선은 두 변수의 관계를 정확하게 묘사하고 있음을 알 수 있다.

표 2.4: 연속형 두 변수의 관계를 파악하기 위한 그래프와 그림함수

그래프	그림함수
산점도 (Scatter plot)	plot
비모수적 회귀모형 (Nonparametric regression)	loess, locfit scatter.smooth
모수적 회귀모형	lm

이렇듯 두 변수의 관계를 매우 유연하게 나타낼 수 있는 비모수적 회귀모형의 장점은 그래픽 기법을 적용하고자 하는 경우에 더 크게 부각됨을 알 수 있다. 그러나 비모수적 회귀도 학부수준에서는 거의 강의가 되지 않고 있는 것으로 알고 있으며 따라서 기본적인 내용에 대한 소개는 반드시 이루어져야 한다고 본다. 특히 비모수적 회귀모형의 몇가지 방법 중에서 가장 좋은 특성을 지니고 있는 국소다항회귀모형은 다른 그래프에서도 많이 사용되고 있기 때문에 반드시 강의내용에 포함시켜야 할 것이다. 국소다항회귀모형에 대한 자

세한 내용은 Fan과 Gijbels (1996)에서 찾아볼 수 있다.

2.3. 삼변량 자료를 위한 그래픽스

세 변수의 관계를 알아보기 위한 첫 번째 작업 역시 산점도를 그리는 것이다. 그러나 2차원 평면인 컴퓨터 모니터 등에 그려지는 세 변수의 3차원 산점도는 입체감을 전혀 느낄 수 없기 때문에 세 변수의 관계를 제대로 파악하는 데에 큰 도움이 되지 않는다. 따라서 2차원 공간에서 세 변수의 산점도를 그리는 방법이 더 효율적이라고 하겠는데, 표 2.5에 소개되어 있는 산점도 행렬과 조건부 산점도가 많이 사용되고 있다.

표 2.5: 연속형 세 변수의 관계를 파악하기 위한 그래프와 그림함수

그래프	그림함수
산점도 행렬 (Scatter plot matrix)	pairs
조건부 산점도 (Conditioning plot)	coplot

산점도 행렬과 조건부 산점도는 각기 다른 장단점을 지니고 있다. 따라서 세 변수의 관계를 제대로 파악하기 위해서는 두 그래프를 모두 그려야 할 것이다. 그림 2.4와 그림 2.5는 자료 ethanol의 반응변수 NOx와 두 설명변수 C와 E의 산점도 행렬과 조건부 산점도이다. 데이터 프레임 ethanol은 패키지 lattice에 있는 자료이며 자세한 설명은 Cleveland (1993)에서 찾아 볼 수 있다.

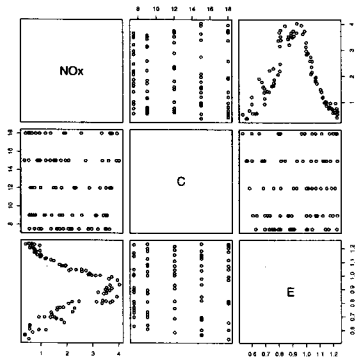

그림 2.4: 세 변수의 산점도 행렬

그림 2.4의 산점도 행렬에는 변수 NOx와 C 사이에 아무런 관계가 없는 것으로 나타나 있으나, 변수 E를 조건변수로 하는 조건부 산점도인 그림 2.5에는 변수 NOx와 C의 선형관계가 잘 나타나 있다. 또한 변수 C와 E 사이에 상호작용이 존재한다는 것도 조건부 산점도를 통하여 확인할 수 있다.

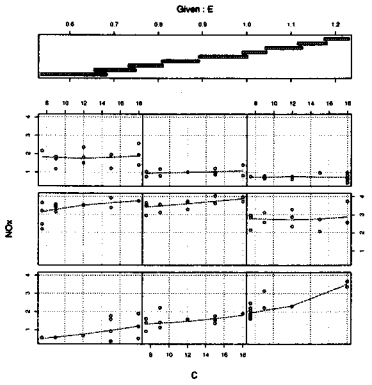

그림 2.5: 세 변수의 조건부 산점도

표 2.6: 삼변량 자료에 대한 반응표면을 그리기 위한 그래프와 그림함수

그래프	그림함수
등고선 그래프 (Contour plot)	contour
색깔을 이용한 등고선 그래프	filled.contour
투시도 (Perspective view)	persp

삼변량 자료에 대하여 매우 유용한 정보를 주는 또 다른 그래프는 반응표면을 나타내는 것이다. 등고선 그래프와 투시도로 나타내는 반응표면의 그래프는 또 다른 측면에서 세 변수의 관계를 파악할 수 있는 기회를 제공해 준다.

투시도는 그래프를 바라보는 각도에 따라서 모습이 많이 변할 수 있다. 따라서 정확한 정보를 얻기 위해서는 가능한 다양한 각도에서 투시도를 그려봐야 할 것이다. 그림 2.6은 자료 ethanol의 세 변수에 대한 반응표면을 비모수 회귀모형으로 추정하고 그 결과를 투시도로 나타낸 것이다.

2.4. 초변량 자료를 위한 그래픽스

초변량 자료는 변수의 개수가 4개 이상인 자료를 의미한다. 실제 자료를 분석할 때 우리가 접하게 되는 대부분의 경우라 하겠다. 그러나 4차원 이상의 공간은 우리가 인지할 수 없는 공간이고, 따라서 초변량 자료만을 위한 특별한 그래픽 기법이 존재하지는 않는다. 다만 주어진 전체 변수 중 일부 변수를 선택하여 그래프를 그릴 수밖에 없기 때문에 자료 전체의 모습을 파악하기 위해서는 수 많은 그래프를 그려야 할 것이다. 그러므로 변수의 개수가 많은 자료를 그래픽스로 분석한다는 것은 곧 많은 시간과 노력을 필요로 한다는 것을 의미한다고 하겠다. 그러나 시간이 많이 소요된다는 이유 때문에 그래픽스에 의한 심층적인 분석을 생략하는 것은 마치 시각 장애인이 코끼리를 만져서 생김새를 추측하는 것과 다

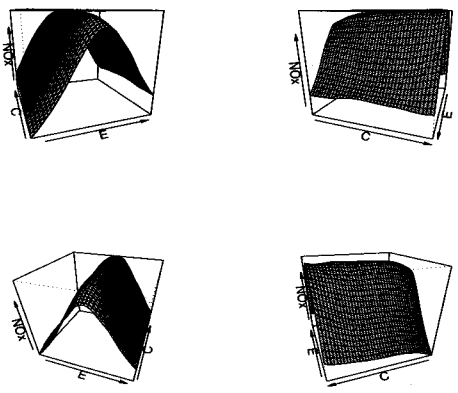

그림 2.6: 자료 ethanol의 세 변수에 대한 반응표면을 투시도로 나타낸 그래프

름이 없다는 것을 학생들에게 강조해야 할 것이다.

그림 2.7과 그림 2.8은 그래프만으로 판별분석을 실시한 예를 보여주는 것으로 학생들에게 그래픽 기법의 가능성을 인식시켜 줄 수 있을 것으로 본다. 자료 iris는 세 종류의 붓꽃에 대하여 꽃받침 조각의 길이와 폭, 꽃잎의 길이와 폭을 각각 측정한 자료로서 여러 문헌에서 자주 인용되고 있는 유명한 자료이다. 분석의 목적은 측정된 네 변수를 이용하여 붓꽃의 종류를 판별해 낼 수 있는 방법을 찾아내는 것이다. 그림 2.7은 네 변수의 산점도 행렬이다.

꽃잎의 길이와 폭을 나타내는 Petal.Length와 Petal.Width의 산점도에서 세 종류의 붓꽃이 명확하게 분리되고 있음을 알 수 있다. 즉, setosa는 두 변수가 동시에 작은 값을 갖는 경우이고 virginica는 두 변수가 동시에 큰 경우이며 versicolor는 그 중간에 분포되고 있음을 알 수 있다. 따라서 두 변수의 곱을 판별변수로 이용하는 것이 효과적이라고 보여진다. 그림 2.8은 두 변수의 곱으로 붓꽃의 종류를 판별한 결과를 나타내고 있다. 전체 150송이 중 오직 네 송이의 붓꽃만이 잘못 분류되었는데, 이것은 이 분류방법이 적어도 주어진 자료에 대해서는 매우 효과적인 분류방법임을 입증하는 것이라고 하겠다.

2.5. Trellis 그래픽스

이차원 이상의 자료를 효과적으로 시각화하기 위해서는 여러 개의 그래프를 잘 정렬해서 배치해야 하는데, 이러한 작업이 R에서 가능하다는 것도 큰 장점 중에 하나라고 하겠다. 그러나 이러한 장점들을 실질적으로 잘 살리기 위해서는 그래픽 모수 등의 활용에 익숙해져야 할 것이다. Trellis 그래픽스는 이차원 이상의 자료를 효과적으로 시각화하기 위하여 Cleveland가 개발한 방법이다. 여러 개의 패널이 행과 열의 구조로 정렬되어 있는 것이 일반적인 모습으로, 비교적 간단하게 여러 개의 그래프를 잘 정렬해서 배치할 수 있다. 그림

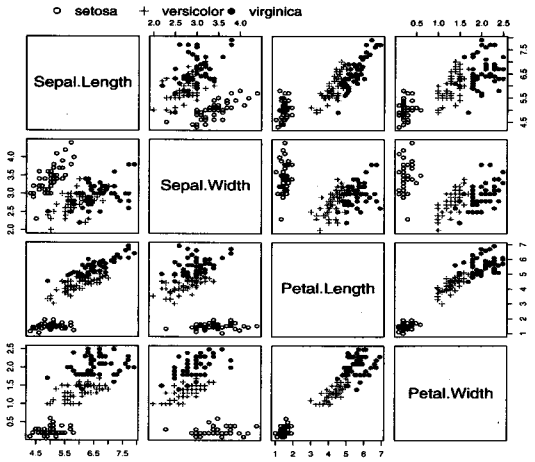

그림 2.7: 자료 iris의 산점도 행렬

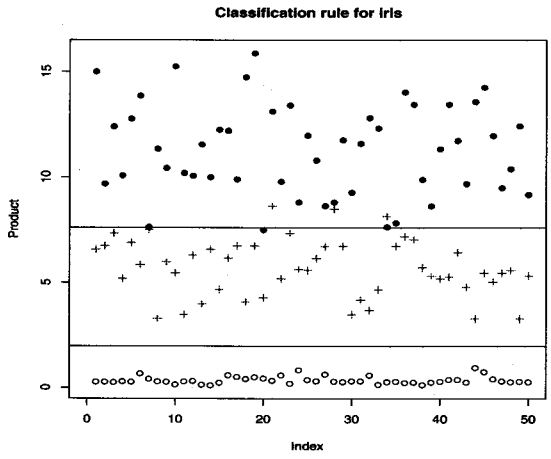

그림 2.8: 자료 iris에 대한 분류법칙 및 결과

1.1이 바로 Trellis 그림함수로 그린 그래프로서, Trellis 그래픽스의 특징을 잘 살펴볼 수 있다.

Trellis 그래픽스에서도 물론 단일 패널에 그래프를 그릴 수 있으며 2.4절까지에서 소개되었던 그래프와 동일한 성격의 그래프를 모두 그릴 수 있으나, 따로 구분하여 학생들에게 소개하는 것이 더 효과적일 것이라고 판단되는데, 그것은 2.4절까지에서 소개되었던 일반적인 그림함수와 다른 점들이 있기 때문이다. 우선 일반적인 그림함수의 그래픽 모수를 조

절하는데 사용되는 함수 par()가 Trellis 그림함수에는 아무런 영향을 미치지 못한다는 점이다. 예를 들어 일반적인 그림함수의 경우에 한 페이지에 두 개의 그래프를 같이 그리기 위해서는 par(mfrow=c(2,1))을 먼저 실행시키면 되는데, Trellis 그림함수의 경우에는 인자 position(혹은 split)과 more를 그림함수와 함께 사용해야 한다. 따라서 약간 더 번거롭기는 하지만 두 개 그래프의 크기를 마음대로 조절할 수 있다는 장점이 있다.

또 다른 차이점은 그래프에 점이나 선 등을 추가하는 방법이 다르다는 것이다. 다양한 그래프를 R에서 쉽게 그릴 수 있는 이유가 높은-수준의 그림함수로 그려진 새로운 그래프에 낮은-수준의 그림함수로 점이나 선 등을 쉽게 추가할 수 있기 때문인데, 일반적인 그림함수의 경우에 함수 plot이나 dotchart, hist 등의 높은-수준의 그림함수로 그래프를 그린 후에 lines나 points 등의 낮은-수준의 그림함수로 원하는 내용을 차례로 추가할 수 있다. 그러나 Trellis 그래픽스의 경우에는 그래프가 여러 페이지에 걸쳐서 출력되는 것을 가정하고 있기 때문에 일단 그래프를 그리고 나중에 점이나 선 등을 추가하는 작업이 불가능하게 된다. 따라서 이러한 작업들을 한꺼번에 수행해야 하는데, 이것은 Trellis 그래픽스에서 실질적으로 그래프를 그리는 역할을 담당하고 있는 것이 그림함수 자체가 아니라 패널함수이기 때문에 가능하게 된다. 즉, 여러 개의 패널함수들을 그림함수에 포함시킴으로 해서 다양한 그래프를 그릴 수 있게 된다.

Trellis 그래픽스는 특히 복잡한 구조를 지니고 있는 자료의 시각화에 탁월한 성능을 갖

표 2.7: 각 그래프를 그리기 위한 Trellis 그림함수와 일반적인 그림함수

자료 유형	Trellis 그림함수	그래프 형태	일반적인 그림함수
일변량 자료	barchart()	막대그림표	barplot()
	dotplot()	점그림표	dotchart()
	bwplot()	상자그림	boxplot()
	stripplot()	점그림	stripchart()
	histogram()	히스토그램	hist()
	densityplot()	커널밀도함수 추정	plot(density())
	qqmath()	분위수-분위수 그림	qqnorm()
이변량 자료	bwplot()	나란히-서-있는 상자그림	boxplot(), plot()
	stripplot()	다중 점그림	stripchart()
	xyplot()	산점도	plot()
	qq()	분위수-분위수 그림	qqplot()
삼변량 및 초변량 자료	contourplot()	등고선 그래프	contour()
	levelplot()	색을 이용한 등고선 그래프	filled.contour()
	wireframe()	투시도	persp()
	cloud()	3차원 산점도	없음
	splom()	산점도 행렬	pairs()
	parallel()	평행좌표 그래프	없음

고 있는 방법이기 때문에 학생들에게 자세한 내용을 소개할 필요가 있다고 본다. Trellis 그림함수들은 패키지 lattice에 있으며, 각 자료의 유형별로 많이 사용되고 있는 주요한 그래프를 그리기 위한 Trellis 그림함수의 목록이 표 2.7에 있다. Trellis 그림함수는 그래프를 그리는 방식에서 일반적인 그림함수와 적지 않은 차이점이 있기 때문에 학생들이 많은 어려움을 호소하는 분야이기도 하다. 따라서 자세한 설명과 충분한 실습이 있어야 할 것이다.

2.6. 적절한 참고문헌의 소개

R과 관련된 수많은 문헌들이 국내외에서 출간되었고, R에 의하여 구현되는 통계 그래픽스에 관련된 문헌도 최근 다수 출간되었다. 교재 내지는 참고문헌으로 사용하기에 적합한 몇몇 문헌들에 대한 간단한 소개가 필요하다고 생각된다.

우선 R과 직접적인 관련을 갖고 있지는 않지만 통계 그래픽스에서 매우 중요한 위치를 점하고 있는 문헌으로 Cleveland (1993)를 들 수 있다. 이 책은 그래픽스가 자료분석과정에서 매우 중요한 도구로 쓰일 수 있음을 보여주고 있으며, 실제로 이 논문에서 제안하고 있는 교과내용의 근간을 제공해 주고 있다. Tufte (2001)는 통계 그래픽스의 역사서라고 할 수 있는 책으로 역사적으로 매우 중요한 의미를 지니고 있는 많은 그래프들을 소개하고 있다. 또한 그래프를 어떻게 설계하는 것이 가장 효과적으로 정보를 전달할 수 있는지에 대해서도 소개하고 있다. Unwin 등 (2006)은 자료의 크기가 점점 커지고 있는 현 상황에 맞추어 대규모의 자료에 적합한 그래픽스의 사용법을 소개하고 있다.

R에 의하여 구현되는 그래픽스의 자세한 소개서로는 단연 Murrell (2006)이 가장 뛰어나다고 하겠다. 이 책의 주안점은 R 그래픽 시스템의 소개이다. 따라서 어떤 그래프가 주어진 자료에 가장 적합한지 등에 대한 논의는 생략되어있다. Maindonald와 Braun (2003)도 눈 여겨 볼만한 책이다. 다만 이 책에서는 자료분석기법 전반을 다루면서 각 기법에서 그래픽 기법이 어떻게 적용되는지에 대하여 소개하고 있기 때문에 통계 그래픽스 자체에 대해서는 약간 아쉬운 점이 있다고 하겠다.

끝으로 이 논문에서 그려진 많은 그래프를 R에서 그리기 위한 구체적인 명령문들은 박동련 (2006)에서 찾아 볼 수 있다. 이 논문의 주된 관점을 적절한 교과내용의 제안으로 한정했기 때문에 그래프를 실제로 그리는 방법들은 모두 생략을 하였다.

3. 강의 사례를 통한 강의방법의 논의

통계그래픽스는 철처히 실습위주로 진행하는 것이 효과적인 강의라고 할 수 있다. 따라서 일반 강의실에서 진행되는 다른 전공과목과는 진행방법과 평가방법 등에서 차이가 있어야 한다. 어떤 강의방식이 가장 효율적인지에 대한 고민이 필요하다고 하겠다.

3.1. 강의 사례 소개

한신대학교 정보통계학과에서는 1999학년도부터 3학년 전공선택 과목으로 통계그래픽스를 개설하여 강의해 오고 있다. 1999학년도부터 2004학년도까지는 S-Plus를 사용하였고,

2005학년도부터 R을 사용하고 있는데, 매년 강의방식과 내용에서 약간의 차이가 있기 때문에 2006학년도 강의를 위주로하여 강의 사례를 소개하고자 한다.

우선 수강생은 34명이었다. 인기과목이라고 할 수는 없겠지만 학생들이 기피하는 과목은 아니라고 하겠다. 수강생들은 모두 R을 처음 접하는 학생들이었고, 강의는 1인당 1PC의 사용이 가능한 전산 강의실에서 3시간 연속강의로 진행되었다. 3시간 중 약 2시간 반정도는 파워포인트로 작성된 강의 내용을 빔프로젝터를 이용하여 학생들에게 설명하고 학생들은 제시된 내용에 맞추어서 실습을 하였으며, 나머지 30분 동안은 학생들이 지정된 연습문제를 풀어서 제출하도록 하였다. 평가는 중간과 기말에 치루는 두 번의 시험결과와 매 강의시간마다 제출하는 연습문제의 평가결과를 각각 동일한 비중으로하여 이루어졌다. 시험은 일반 강의실에서 open book으로 답안지에 명령문을 적는 방식으로 치루었다.

3.2. 학생들의 평가내용

학기가 끝난 후 학생들에게 설문조사를 실시하여 R로 구현하는 통계 그래픽스 과목을 수강하면서 좋았던 점과 어려웠던 점 등을 조사하였다. 다음은 주요 내용이다.

1. 좋았던 점 및 강의를 듣고 도움이 된 점

   - 다른 통계 프로그램에 비해 수식 계산이 쉽다.
   - 대략적인 data screen 작업이 신속하고 편리하다.
   - 자신의 생각을 그래프로 표현하는 것이 가능하다.
   - 무료 소프트웨어이기 때문에 어디서나 편리하게 사용할 수 있다.
   - 색상이 다양해서 그래픽 표현에 적합하다.
   - 손으로 그래프를 그리는 것 같다는 느낌을 받는다.
   - 다른 과목의 발표에서 그래픽을 유용하게 이용할 수 있었다.
   - 그래프 해석능력이 향상되었다.
   - 새로운 그래프를 접할 수 있어 좋았다.

2. 어려웠던 점 및 개선이 필요한 점

   - 프로그램을 수행 할 때 에러가 많이 발생하며 에러발생 원인의 파악이 어렵다.
   - 프로그램의 기본지식을 가지고 있어야 한다. 즉, C 언어와 같은 다른 언어를 하나 정도 꼭 알고 있어야 할 것 같다.
   - SAS를 처음 접할 때와는 또 다른 어려움을 느낀다.
   - 반을 두 반으로 나누어 수준을 다르게 강의했으면 좋겠다. 수준차이가 너무 나서 실습시 대기시간이 너무 길다.
   - 1주일에 한 번의 강의로는 R을 이해하는데 어려움이 있다. 지속적인 훈련이 가능하도록 과제물이 있어야 할 것 같다.

- 더 정확한 평가를 위해서 시험은 컴퓨터를 이용해서 보는 것이 좋을 것 같다.

전체적으로 R이 그래픽에 탁월한 성능을 지녔다라는 점에는 모두 수긍을 하고 있으나, SAS와 같은 다른 통계 프로그램에 비해 아직도 사용하기 어렵고 불편하다는 의견이 많았다. 또한 앞으로 그래프를 그리기 위해 R을 사용할 것인가라는 질문에는 약 절반정도의 인원만이 그렇다라는 대답을 하여 한 학기동안 R을 다루었지만 아직 많은 학생들이 어려움을 느끼고 있으며 조금은 기피하는 경향이 있음을 알 수 있었다.

3.3. 문제점 및 대안

가장 큰 문제는 학생들이 R을 처음 접했다는 것이다. 물론 R을 처음부터 능숙하게 다룰 수 있어야 하는 것은 아니지만 처음 접하는 소프트웨어를 이용하여 한 학기안에 그래프를 편하게 그릴 수 있는 수준에 도달하는 것은 학생들에게 무척 어려운 일임에 틀림이 없으며, 따라서 미리 R을 접할 수 있는 기회를 제공하는 것이 꼭 필요하다고 하겠다. R이 갖고 있는 장점을 고려해 볼 때 가능한 빨리 학생들에게 R을 소개하는 것이 좋다고 판단되어 한신대학교에서는 2007학년도부터 1학년 전공필수 과목인 정보통계학개론에서 R을 소개하고 있다. 또한 다른 전공과목에서도 SAS만을 강조하는 것이 아니라 R을 함께 소개할 예정에 있다.

강의가 진행됨에 따라서 학생들 사이에는 아무래도 성취도 차이가 날 수 밖에 없게 된다. 주어진 연습문제 등을 해결하는데 있어서 못하는 학생들은 프로그램에 자주 오류가 발생하게 되며 그러한 오류의 수정이 거의 불가능한 상태가 되어 담당교수의 도움이 절대적으로 필요로 하게 되는 반면, 잘하는 학생들은 상대적으로 빠른 시간안에 해결할 수 있기 때문에 결국 긴 대기시간을 가질 수 밖에 없다는 문제가 생기게 된다. 이러한 학생들의 수준차이로 인한 강의진행 속도의 조절문제는 모든 실습과목에서 발생하는 문제라고 하겠다. 해결방안으로는 실습조교의 활용이 있을 수 있으며, 빨리 끝내고 기다리는 학생들에게 따로 심화문제를 더 주는 방법도 생각해 볼 수 있다. 또한 학생들의 프로그램 오류를 빨리 수정해 주기 위해서는 R 전반에 대한 명확한 지식이 담당교수에게 요구된다고 하겠다.

학생들의 성취도 평가방법도 많은 고민이 필요한 부분이다. 중간고사와 기말고사를 볼 때 컴퓨터에서 R을 실행시키면서 답안을 작성하도록 하는 것이 바람직한 평가방법이라고 생각되지만, 전산실습실에서는 옆 사람과 충분한 거리를 확보할 수 없기 때문에 부정행위의 가능성이 더 높아진다는 문제가 있다. 또한 학생들이 제출하는 과제물도 혼자서 해결했는지 여부를 확인할 수 있는 방법이 마땅치 않다는 문제가 있다. 정확하면서도 공정한 평가방법을 고안해야 할 것이다.

몇 가지 어려움이 있는 것은 사실이지만 통계 그래픽스 과목은 학생들에게 매우 유용한 과목임에는 틀림이 없다. 그러나 이 과목을 이수했다고 하더라도 자료분석 도구로서 그래픽 기법을 자유자재로 활용하기에는 많은 어려움이 있는 것 또한 사실이라고 하겠다. 따라서 다른 전공과목에서도 R로 구현되는 그래픽 기법을 적극 활용하도록 하는 것이 절대적으로 필요하다고 본다.

참고문헌

박동련 (2006). <R에 의한 통계그래픽스>, 자유아카데미.

Anscombe, F. J. (1981). *Computing in Statistical Science through APL*, Springer-Verlag, New York.

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the old faithful geyser, *Applied Statistics*, **39**, 357–365.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press, Summit, New Jersey.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, John Wiley & Sons, New York.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman & Hall/CRC, London.

Fisher, R. A. (1971). *The Design of Experiments*, 9th ed., Hafner, New York.

Maindonald, J. and Braun, J. (2003). *Data Analysis and Graphics Using R - an Example-based Approach*, Cambridge University Press, New York.

Murrell, P. (2006). *R Graphics*, Chapman & Hall/CRC, Boca Raton, Florida.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer-Verlag, New York.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*, 2nd ed., Graphics Press, Cheshire, Connecticut.

Unwin, A., Theus, M. and Hofmann, H. (2006). *Graphics of Large Datasets*, Springer-Verlag, New York.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*, Chapman & Hall/CRC, New York.

[2007년 6월 접수, 2007년 7월 채택]

Teaching Statistical Graphics using R

Dongryeon Park[1)]

ABSTRACT

It is well known that graphical display is critical to data analysis. A lot of research for data visualization has been done, so many effective graphical tools are now available. With the proper use of these graphical tools, we can penetrate the complex structure of data set easily.

To enjoy the benefit of the powerful graphical display, the choice of the statistical software is very crucial. R is a popular open source software tool for statistical analysis and graphics, and can provide the very powerful graphics facilities. Moreover, many researchers believe that R is the best software for statistical graphics.

In this paper, we would like to discuss what we teach and how we teach in statistical graphics course using R.

Keywords: Data analysis tool, statistical graphics, R.

1) Professor, Department of Statistics, Hanshin University, Osan, Kyunggi-do 447-791, Korea
E-mail: drpark@hs.ac.kr