디지털 사회
DIGITAL SOCIETY
제2호

연세대학교 사회과학데이터혁신연구센터
SOCIAL SCIENCE DATA INNOVATION CENTER

발 행 인: 조화순
발 행 일: 2017년 10월 2일
홈페이지: http://democracy3.0.yonsei.ac.kr

데이터 분석 도구로서의 R언어의 장점과 단점

이병재
(연세대 사회과학데이터혁신연구센터 연구교수)

데이터는 사실 모든 곳에 존재한다. 데이터(data)라는 단어는 라틴어 동사의 “give”를 의미하는 동사인 “dare”에서 유래해서 “something given”을 의미하는 라틴어의 datum의 복수형이다. 예전에는 철학에서 소여(所與)라고 번역되기도 했으며, 사유에 의해 가공되지 않은 직접적인 의식 내용을 의미했다. 일반적으로 분석을 거치지 않은 자료를 의미한다고 하겠다. 모든 종류의 실증연구에서 데이터의 역할은 매우 중요해서 에드워드 데밍(Edward Deming)이 말한 “In God We Trust, All Others Bring Data.”라는 언급은 많은 실증연구 분야의 모토로 여겨지고 있다. 그러면, 사회과학 분야에서 사용되는 데이터는 무엇인가?

불과 20년 전만 하여도 사회과학분야에서 사용되는 데이터는 주로 정부기관이나 국제기구가 수집한 양적 데이터, 서베이 데이터, 참여관찰 등의 방법으로 수집된 질적 데이터에 한정되어 있었다. 하지만 인터넷의 발달을 비롯하여 인간생활의 모든 면이 혁명적인 변화를 겪으면서, 생산되는 데이터의 양과 종류 역시 급격하게 증가하고 다양화되어, 지난 1년 동안 만들어진 데이터의 양이 인류가 그 이전에 만들어낸 데이터의 양보다 더 많다는 말 자체가 상투적인 표현으로 여겨지고 있다. 흔히 말하기를, 현재 디지털 세계에 2.7 제타바이트의 데이터가 존재하고, 2025년에는 180제타바이트가 존재할 것이라고 예상된다. 제타바이트(zettabyte)는 1000의 7제곱, 즉 10의 21제곱을 말한다. 문제는 엄청난 속도로 쏟아져 나오는 이 데이터에 대한 우리의 분석능력이 생산되는 데이터의 양을 따라가지 못한다는 점이다. 어떤 의미에서 분석되지 않은 데이터는 위에서 말한 대로 단순히 주어진 것일 뿐이다.

데이터 과학은 이러한 상황에서 데이터를 분석하기 위해 새롭게 등장한 분야이며, 우리가 이전에 익숙하게 사용하던 양적, 질적 데이터 수집 및 분석을 위해 사용하던 도구들은 물론 전대미문의 규모와 다양성에 걸맞는 새로운 수집 및 도구들을 필요로 한다. 일반적으로 데이터 과학에서는 통계, 수학, 프로그래밍, 문제해결, 데이터 획득 등을 창조적인 방식으로 결합할 능력이 요구되는데, 이 과정에서 데이터 정제라 불리는 과정을 비롯한 데이터 준비 및 데이터 결합을 물론 데이터에서 다양한 종류의 패턴을 인식하는 능력이 필요하다. 왜냐하면 데이터 과학은 일견 무질서하게 산재하는 데이터를 정제하고, 준비하고, 분석하는 과정이기 때문이다. 간단히 말하면 데이터 과학은 데이터로부터 정보와 지식을 얻어내는 데 사용되는 도구들을 지칭하는 통칭개념이라 할 수 있다. 그렇다면 데이터 분석에 필요한 실제 도구들에는 어떤 것들이 있을까?

데이터 과학자들은 일반적으로 데이터 과학에 필요한 도구들로 다음을 들고 있다. 1. R, SAS, Stata 등의 대한 통계 및 통계분석도구에 대한 깊은 지식 및 숙달, 2. Python, Java, Perl, C++, Julia 등의 코딩 언어에 대한 지식 및 숙달, 3. Hadoop, Hive, Pig, 그리고 SQL 등의 데이터베이스 등에 대한 지식, 4. 구조화되지 않은 데이터 – 소셜 미디어, 비디오 피드, 오디오, 이미지 등 –를 수집 및 분석할 수 있는 능력이다. 물론 한 사람의 데이터 과학자가 이 모든 도구들을 같은 정도로 사용하지는 않으며, 필요에 따라 선택적으로 습득 및 사용한다. 이 글에서는 이 다양한 도구 중에서 최근들어 각광을 받고 있는 R에 대한 간단한 소개를 하고자 한다.

R의 가장 큰 장점은 무료라는 점이다. R은 www.r-project.org에서 무료로 내려받을 수 있다. 소프트웨어 구입을 위한 연구비가 충분하지 않은 연구자들은 물론 다른 연구자들과 공동작업을 할 경우나 교육용으로 통계 프로그램을 사용할 때 무료라는 점은 커다란 장점으로 다가온다. 일반적으로 R은 처음에 배우기 어렵다고 알려져 있으며, 이는 다른 메뉴방식의 통계프로그램 (SPSS나 Stata 등)과 비교할 때 어느 정도 사실이다. 하지만, 예전과 달리, Emacs+ESS, RWinEdt, RVim, Tinn-R 등을 비롯하여 최근의 RStudio라는 인터페이스의 등장으로 R의 사용은 획기적으로 수월해졌으며, 이러한 인터페이스의 발달은 R을 교육용으로 사용하는데 커다란 장애물을 제거했다고 알려져 있다. 또한 R은 대부분의 주요 운영체계에서 작동한다. 이는 사실 윈도우 사용자에게는 특별히 잇점은 아니며, 여타의 통계패키지도 주요 운영체제에서 작동하기는 한다. 하지만, 매킨토시나 리눅스 사용자에게 커다란 장점이다.

둘째, R은 무료이지만 저급의 프로그램이 아니며, 대부분의 통계학자들이 사용하는 프로그램이다. 이는 새로운, 혹은 더나은 기법이 개발되었을 때, 처음으로 구현될 가능성이 높다는 의미이다. 또한 새로운 통계 기법을 사용할 경우 그것을 가능하게 하는 도구나 도움을 줄 수 있는 사람이 존재할 가능성이 훨씬 높다는 의미이기도 하다. 이러한 점은 머신러닝 등을 비롯하여 첨단분야의 작업을 하는 사람에게 매우 중요하다.

셋째, R의 가장 큰 장점은 방대한 패키지 생태계이다. 2017년 현재 14,000개 이상의 R 패키지가 존재한다. R의 확장가능성과 개발자들이 기존의 패키지를 수정하여 데이터 분석을 위한 도구를 스스로 변형 및 개발시킬 수 있다는 것은 다른 상업용 프로그램에서는 발견할 수 없거나 제한적으로만 발견되는 커다란 장점이다. 이러한 장점 때문에, R은 점차로 사회과학은 물론 바이오과학, 인문학(텍스트 데이타 분석, 네트워크 데이터 분석) 등의 분야에서도 사용자가 늘어나고 있다.

넷째, R은 매우 훌륭한 그래픽을 제공한다. 일반적으로 Matlab, Mathematica, Gauss 등이 뛰어난 그래픽을 제공했었지만 R은 이들보다 고급의 시각화도구를 제공한다. 최근에 개발된 Hadley Wickam의 ggplot2은 상상을 뛰어넘은 수준의 훌륭한 시각화를 구현할 수 있다.

다섯째, 기초 통계교육용 목적으로 R은 훌륭한 도구로 사용될 수 있다. 터미널에 실제로 코드를 입력하고 실행시키는 방식은 처음에는 익숙하지 않아서 어렵게 생각되지만, 이러한 과정을 통해서 메뉴방식 프로그램에서와는 달리 실제 작동원리를 이해하는데 훨씬 도움이 된다, 또한 R이 제공하는 다양한 다양한 시뮬레이션 관련 도구들과 사용자 정의 함수 작성의 수월성은 확률 및 통계의 기초개념을 이해하는데 많은 도움을 준다.

여섯째, R은 C라든가 Fortran으로 작성된 프로그램과 무리없이 연계될 수 있다. 이는 보다 처리속도가 빠른 언어로 작성된 프로그램이 가지는 장점을 R에서 이용할 수 있다는 점이다. 이 밖에도 R은 많은 장점을 가지고 있으며, 위에서 말한 바와 같이 많은 뛰어난 연구자들이 R을 사용하고 있기 때문에 그 기능은 하루가 다르게 개선 및 개발되고 있다.

물론 R이 장점만 있는 것은 아니다. 많은 사람들이 R의 단점으로서 메모리 관리, 속도, 효율성을 지적한다. 이 세가지 측면에서 급격한 개선이 이루어지고 있기는 하지만 다른 언어에 비해 부족하다고 지적되고 있다. 부분적으로 R이 1960년대에 만들어진 S언어에 기반하여 만들어진 프로그램이라는 점에 기인한다. 즉 기본적으로 낡은 기술에 기반하여 만들어진 시스템이라는 것이다. 이러한 점은 때때로 대규모의 데이터를 취급할 때 문제를 야기하기도 한다. 기본적으로 데이터는 실제 메모리에 저장되어야 하는데, 이는 취급할 수 있는 데이터의 양을 제한한다. 하지만, 컴퓨터 메모리의 급격한 증가로 인해 이는 사실 그다지 심각한 문제는 아니다.

또 하나의 단점으로 지적되는 것은 보안문제이다. 예를 들어, R을 웹브라우저를 이용하여 원거리에 있는 서버를 이용하여 계산하여 결과를 전송하는 것은 보안문제 때문에 매우 어렵다고 알려져 있다. 하지만, 최근들어 아마존 등에서 제공하는 (Amazon Web Services) 가상 컨테이너 등을 사용하여 이 문제의 심각성은 차츰 감소되고 있다고 한다.

요약하자면, R은 프로그래머들만을 위한 언어가 아니라 데이터를 이용하여 문제해결을 하려는 사람들에게 유용한 컴퓨터 환경이다. 물론, 위에서 언급한 작업들이 R이 아닌 다른 언어 – 예컨대 Python –을 통해서 할 수 없는 것은 아니며, R만이 그러한 작업을 할 수 있는 것도 아니다. 하지만, 데이터 과학을 위해 필요한 도구를 습득해야 한다면 R은 충분히 고려할 만한 가지가 있는 언어라고 하겠다.