Commentary

統計解析環境R言語の紹介

中村 健太郎（埼玉学園大学経営学部専任講師）

共通語としてのR

「いざ我等降り，彼処にて彼等の言葉を乱し，互いに言葉を通ずることを得ざらしめん。」

押井守監督による1989年のアニメーション『機動警察パトレイバー 劇場版』では，旧約聖書創世記11章から上記の語句が引用されます。筆者が大学院生であった2003年頃，テレビで観た『パトレイバー』の「バベルの塔」に関する上述のシーンは，最新の基本ソフトウェアを導入した産業用ロボットが暴走する事件を描くこの映画の中でも，ことさらに強く印象に残りました。

というのも，その当時，統計解析用のソフトウェアやプログラミング言語としてSASやSPSS，MATLAB，MATHEMATICAなどに囲まれ，講義によってはC言語などを使い，特定の統計モデルに関しては専用のソフトウェアの使用法をそれぞれ学ぶ必要がある中で，先行研究の手法がSという言語で実装されていたため，さらに新しい言語を学ぶ必要に迫られていたからです。

教科書の分析例や，統計のそれぞれの手法をバラバラの統計ソフト，言語で理解するのは困難が伴いました。また，効率も悪く，拙い自分の乱れたプログラムでは分析が停止したり，ひどい時には暴走したりしてしまうことも少なくありませんでした。

ところが，近年，計量的な研究の領域やデータ解析の場面で，1つのアルファベットの文字に特に強い関心が集まっています。その文字とは，R，統計解析環境R言語です。Rは現在，*lingua franca*（共通語）という表現が用いられるほど普及し，分析に活用され，さらに発展を続けています。

たとえば，Aitkin et al.（2009）では，それまでGLIM4というソフトウェアの使用を前提にしていた統計モデルの教科書が，R版に書き換えられるという変化が起こっています。その理由として，多くの大学でRの採用が進んでいること，Rは統計的手法を包括的に扱えることなどが挙げられています。竹村（2007）や豊田編（2008）も，改訂にあたってRの使用を想定するようになりました。また，統計数理研究所の公開講座ではRを実習に活用する講義が多くなっています。

さらに，アメリカの新聞*New York Times*紙は，2009年1月6日付の記事においてRの産学双方での広がりを伝えています（Vance，2009a）。記事では，Googleや製薬会社のPfizer，Bank of Americaといった金融，Shellのようなエネルギー産業まで幅広い業種でRが利用されている実態が取り上げられ，大学でのRの採用の拡大，大学院修了生のRの習得の実態などが報じられています。この記事自体に対する反響も大きく，記事発表の2日後には記者がブログで補足を行っており（Vance，2009b），Rに向けられた注目の高さをうかがい知ることができます。

記事によると，Rを主要な統計解析ソフトウェアとして通常使っている利用者は約25万人と推計されています。日本での正確な利用者数を把握することは困難ですが，着実に増加していることは間違いないでしょう。Rに関する，あるいはRの使用を前提とする書籍も多数刊行されており，近年の動向を考えると，Rを知ることは手法の学習や研究，実用などさまざまな場面で効率性，生産性を向上させ，非常に有効であると考えられます。

本稿では，しかしながら，以降でRのインストール方法や，データ解析の実際，具体的なプログラム例などに関する説明は行いません。簡略にRのコードなどを筆者が稚拙に解説することで，Rの魅力を減じてしまうかわりに，R利用の広が

りを伝え，Rの特徴を記述し，現在，多数出版されている優れた教科書のごく一部を紹介することを通じて，Rの導入への補助線となれば幸いです。

Rとは何か

Zuur et al.（2009）は，想像しうるあらゆることをRでは実現可能である，と述べています。Rは効率的なデータ操作，行列の形式にも強い計算機能，統合的かつ豊富なデータ解析手法，さらに，データの分析や表示のためのグラフィクスやプログラミング言語としての機能を備えた汎用的で統合的なソフトウェアです（Venables et al., 2010）。UNIXやMAC OS X，Windowsなど多様なオペレーティングシステム上で使用することができます。

高機能なRは，その一方で，ソースコードを全面的に公開し，自由に複製，配布，改変が可能であるという特徴をもっています。つまり，どのような仕組みで動いているのかを完全に把握できるRを無料で入手し，利用できるということです。このオープンソースとしてのRの特徴は，自発的で，分散的かつ協働的な開発を促し，世界中の人々がRの発展に寄与する状況を生み，現在も改良が続けられています。

Rは更新の頻度が高く，年に2回，4月と10月に細かい機能を追加するなどのマイナーな更新が行われます。それ以外でも，主に小さな不具合の修正であるパッチでの更新が随時行われています。また，R本体だけでなく，分析手法や他のソフトウェアとの連携などさまざまなRの拡張機能が多くの研究者などによって開発され，パッケージという形で入手可能となっています。このパッケージの充実ぶりを考えると，Rではあらゆることが可能であるという上述の意見も，あながち大げさとはいえない印象を受けます。

Rは，統計解析やグラフィクスのための言語であり環境である（R Development Core Team, 2010a）と表現されます。ここで環境（environment）という言葉は，Rが十分に計画された一貫した設計思想のシステムであることを示しています（Venables et al., 2010）。

先述したRのデータ操作や計算，描画などの機能は，有機的に連携させることが可能です。たとえば，回帰分析の出力結果を用いて，今度はそれを図示すること，あるいは次の段階の詳細な分析に利用することなども容易です。回帰分析のRでの実行方法は，わずかな変更で一般化線形モデルにも適用できます。また，要約のための関数を1つ覚えることで，多岐にわたる手法のさまざまな分析結果も，内部での違いをほとんど意識せずに同様に使用することができます。このことは，データの特徴（より細かくは変数の種類など）に応じた適切な処理，分析手法の選択が行われることを意味します。

そもそもRは，「統合的な考え方に基づき，洗練され，広く認められた揺るぎないソフトウェアシステム」として1998年にThe Association for Computing MachineryのSoftware System Awardを受賞したS言語に基づいています。Sはそれまでのデータの扱い方，分析の仕方，視覚化の方法を決定的に変えたといわれる優れたシステムです。そのようなソフトウェアを設計の根幹に置くRによって，統計解析のさまざまな要素を総合的に扱えることは，データを詳細に分析し，分析した結果を使って次の分析に繋げるといった対話的，逐次的，探索的なデータ解析を可能にします。グラフィクスなども併用しながら，データから広く深く知見を導くことができる統一的な操作環境は，定型的で固定的なソフトウェアの出力に飽き足らない分析者にとって，理想的であるといえるでしょう。

しかし一方で，Rの柔軟性は，分析者が何をどうしたいのかという目的を明確にし，何をしているのかという分析の実質を把握する必要性を，より高めることも意味します。基本的には，入力画面に関数と呼ばれる命令を入力することで処理を実行していくRでは，ある程度のプログラミングの知識，技術が必要となります。

視覚的にわかりやすいインターフェイス上で，マウスなどのクリックによって高度な分析も実行可能なソフトウェアに比べると，「学習曲線が急峻である」「敷居が高い」などといわれる所以の1つです。しかし，Rの導入を補助する情報は豊富に用意されています。決して敬遠する必要はありません。

Rの導入

Rはhttp://www.r-project.org/からダウン

ロードすることで入手可能です。実際のダウンロードの際には，CRAN（Comprehensive R Archive Network）と呼ばれるサイトから最寄りのミラーサイト（たとえば兵庫教育大学や筑波大学）を選択してダウンロードすることとなります。

Rのホームページ The R Project for Statistical Computingには，多くの情報が掲載されています。その中でも，7種類から成るマニュアルは非常に詳細です。インストールに関しては，R Development Core Team（2010b）にさまざまなオペレーティングシステムに応じた詳しい解説があります。また，Venables et al.（2010）はRの入門的内容を扱っています。初歩的な導入から網羅的にRの機能，特徴が記述されており，豊富な内容が凝縮されています。

Rは，オンライン上のマニュアルが充実しているだけではなく，ヘルプも非常に詳しいことが特徴です。使い始めの最初のうち，Rの関数を使って処理を実行していく際に，何かわからないことや困ったことがあったら，ヘルプを参照するだけで解決されることが多々あります。Rを使い慣れているつもりでも，ヘルプの関数の説明に新たな発見があることも珍しくありません。

特定の関数の詳細を調べるにはhelp（関数名）と入力します。なお，関数の本体がどうなっているのかを知るには，（ ）を付けずに関数名だけ，たとえばhelpとすれば，関数helpの中身（Rでどのような処理をする関数なのか）が表示されます。

ある関数の詳細を調べるだけでなく，目的の処理に対応する関数を見つける検索のための関数も存在します。指定した語句の全部，または一部を含む関数を探したり（apropos），曖昧なマッチングで検索したり（help.search）することが可能です。また，前述のマニュアルやFAQ（Frequently Asked Questions）を含むRに関する情報がHTML文書化されており，ブラウザを起動してそれらを閲覧する関数（help.start）も重宝します。ここでは，キーワード検索を実行することも可能です。

ある関数が，実際どのように使用できるかのデモンストレーションを実行させたり，具体的な使用例を表示させる関数（demoとexample）は，Rを実際的に理解するのに大変便利です。自分自身で入力し，結果を確認する作業の他に，これらの関数やヘルプを積極的に活用すると，Rで実際に何が可能なのかが，その実現方法とともに一目瞭然となり，理解が早く深まることが期待できます。

Rはメーリングリスト上でも活発に議論されています。Rに関する各種文書に加えて，メーリングリストに蓄積された知識は，RSiteSearchという関数で検索可能です。

WEB上の日本語の情報は，岡田昌史先生が管理されているRjpWikiに膨大かつ詳細に掲載されています（http://www.okada.jp.org/RWiki）。情報は文字通り日々更新され，追加されています。日本語でのインストール方法やRを活用するうえでの有益な情報の他に，質問と回答が寄せられる掲示板でも活発な議論が行われています。

Rの書籍

一般的な統計手法の解説書としては，山田ほか（2008）が大変参考になります。インストール方法から丁寧に説明されており，記述統計，推測統計の基本から因子分析や共分散構造分析などの多変量解析，擬似乱数によるシミュレーションや検定力分析といった独自性の高い重要な内容までが，幅広く，興味深い例題データとともに解説されています。たんにRでの実行例が羅列されるのではなく，理論的な説明が平易になされる一方で，Rによる実例が示されるので，この一冊だけでも統計学の知識とRの技術を深く身に付けられるでしょう。

また，青木（2009）ではRの特徴を活用して効率的にデータを分析するための有効な情報を得ることができます。その中でも特に，データの取り扱い方に関する説明は非常に参考になります。また，書籍のサポートページ（http://aoki2.si.gunma-u.ac.jp/R）の情報も実際のデータ解析で役立つものばかりです。

一方，土屋（2009a）は，社会調査を進めるうえで欠くことのできない標本調査法について，詳しく論じています。本書自体ではRによる実習は記述されませんが，朝倉書店からダウンロードできる付録（土屋，2009b）において，標本調査データの分析をRで実行する手順が詳しく解説されています。Rによる演習問題とその詳しい解答が記載されているので，理論と実習を相補的に

往復することができます。

土屋（2009b）では，Rの機能を拡張する関数やデータのまとまりであるパッケージのうち，surveyというパッケージを用いて，標本抽出デザインに従って収集された調査データの分析を行っています。Lumley（2010）は，このsurveyパッケージの作者による標本抽出法の解説です。本文中にRのコードと出力結果が挿入されています。本文の内容をRで実践的に理解しながら読み進めるのに適しています。

また，星野（2009）は，データ分析に関わる者として知っておくべき「偏りのあるデータ」に対する理論書ですが，傾向スコア解析など注目の手法の実行例を付録においてRで示しています。実践的な問題を考える端緒として大変参考になります。

より具体的にRを調査データ解析に利用する解説書として，緒賀（2010）があります。心理学研究におけるデータ解析を念頭に置いていますが，Rコマンダー（舟尾，2008）と呼ばれるグラフィカルユーザーインターフェースを使用しての分析方法を示しており，まずは手元のデータをRによって分析したいといった場合に参考になる文献です。なお，信頼性や潜在変数モデルの推定といったテスト理論（池田，1994）や計量心理学（岡本，2006）に関する話題は，CRANのPsychometric Models and Methodsの項にまとめられており（http://cran.r-project.org/web/views/Psychometrics.html），使用可能なパッケージなどについて詳述されています。

これに対して，竹内（2005）は，統計解析環境としてRを導入するのではなく，時には雑学やクイズのような問題をRの関数で実装していき，プログラミング言語としてのRに入門することを志向しています。Rに関して一部古くなっている記述もありますが，統計モデルの実習環境としてRを理論や概念とともに学ぶのではなく，Rそのものについてまずは知っておきたい，慣れておきたいという場合に有用です。あるいは，Ligges（2004）やZuure et al.（2009）もR自体の習得を意識したものとなっています。

Rの特徴の1つである高度な描画性能については，たとえばMurrell（2006）やSarker（2008）などに詳述されています。さらに，入門的な文献ではありませんが，Rを使い続けていくうえで間瀬（2007）は大変参考になります。

おわりに

すでに述べたように，Rは無料で利用可能です。しかし，無保証でもあります。この点に関して，Keeling and Pavur （ 2006 ） やAlmiron et al.（2009）は，Rの正確性について実証的に検討しています。Rのすべてについて検証されているわけではありませんが，Rの高い信頼性を示す結果を得ています。

慣れているソフトウェアから離れて，あるいは初めての統計解析ソフトとしてRを使い始めると，操作に戸惑ったり，難しく感じたり，敷居が高いと感じたりするかもしれません。しかし，Rを使用するために必要な資格（ライセンス）に特別なものはいりません。それは関数licenseの実行で表示される言葉に象徴されています。

Share and Enjoy.

データ解析を楽しみながら学べる，研究できる，使える環境こそRなのです。ぜひ実際に使ってみて，その喜びを共有してみてください。

文献

Aitkin, M., B. Francis, J. Hinde and R. Darnell, 2009, *Statistical Modelling in R*, Oxford, U.K.: Oxford University Press.

Almiron, M.G., E.S. Almeida and M.N. Miranda, 2009, "The reliability of statistical functions in four software packages freely used in numerical computation," *Brazilian Journal of Probability and Statistics*, 23(2): 107-19.

青木繁伸，2009，『Rによる統計解析』オーム社。

舟尾暢男，2008，『「R」Commanderハンドブック――A Basic-Statistics GUI for R』オーム社。

星野崇宏，2009，『調査観察データの統計科学――因果推論・選択バイアス・データ融合』岩波書店。

池田央，1994，『現代テスト理論』朝倉書店。

Keeling, K.B. and R.J. Pavur, 2007, "A Comparative Study of the Reliability of Nine Statistical Software Packages," *Computational Statistics & Data Analysis*, 51(8): 3811-31.

Ligges, U., 2004, *Programmieren mit R.*, Springer.（石田基広訳，2006，『Rの基礎とプログラミング技法』シュプリンガー・ジャパン。）

Lumley, T., 2010, *Complex Surveys: A Guide to Analysis Using R.*, Hoboken, N.J.: Wiley.

間瀬茂，2007，『R プログラミングマニュアル』数理工学社。
Murrell, P., 2006, ***R Graphics***, Boca Raton, Fla.: Chapman & Hall.（久保拓弥訳，2009，『R グラフィックス――R で思いどおりのグラフを作図するために』共立出版。）
緒賀郷志，2010，『R による心理・調査データ解析』東京図書。
岡本安晴，2006，『計量心理学――心の科学的表現をめざして』培風館。
R Development Core Team, 2010a, ***R: A Language and Environment for Statistical Computing***, Vienna, Austria: R Foundation for Statistical Computing.
――――, 2010b, ***R Installation and Administration***, Vienna, Austria: R Foundation for Statistical Computing.
Sarkar, D., 2008, ***Lattice: Multivariate Data Visualization with R.***, New York: Springer.（石田基広・石田和枝訳，2009，『R グラフィックス自由自在』シュプリンガー・ジャパン。）
竹村彰通，2007，『統計〔第 2 版〕』共立出版。
竹内俊彦，2005，『はじめての S-PLUS/R 言語プログラミング――例題で学ぶ S-PLUS/R 言語の基本』オーム社。
豊田秀樹編，2008，『データマイニング入門――R で学ぶ最新データ解析』東京図書。
土屋隆裕，2009a，『概説 標本調査法』朝倉書店。
――――, 2009b，『概説 標本調査法』付録（第 1.0 版），朝倉書店。
Vance, A., 2009a, "Data Analysts Captivated by R's Power," ***The New York Times***, January 6.（http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html）.
――――, 2009b, "R You Ready for R?"（http://bits.blogs.nytimes.com/2009/01/08/r-you-ready-for-r/）.
Venables, W.N., D.M.Smith and the R Development Core Team, 2010, ***An Introduction to R.***, Vienna, Austria : R Foundation for Statistical Computing.
山田剛史・杉澤武俊・村井潤一郎，2008，『R によるやさしい統計学』オーム社。
Zuur, A.F., E.N.Ieno and E.H.W.G.Meesters, 2009, ***A Beginner's Guide to R.***, New York: Springer.