統計数理研究所で開発されたRパッケージ

中野　純司　モデリング研究系　教授

【はじめに】

統計数理研究所では計算機の出現以来、その時代の最新の統計計算環境を維持してきた。そしてその上で種々のソフトウエアが開発されてきている。それらの中には40年以上前に作成されたにもかかわらず、現代でもその価値を失っていないものもある。

しかしながら計算機技術は急速に発展しており、過去のソフトウェアをソースコードでそのまま配布しても使える人は限られてしまう。そのため統計科学技術センターを中心として、それらのソフトウェアをその時代の計算機環境に合うように保守し、さらに利用しやすくすることを試みてきた。近年は統計科学分野でデファクトスタンダードになっている統計解析ソフトウェアRからそれらのソフトウェアを利用できるように、Rのパッケージとして CRAN などで広く公開している。

本発表ではそれらのRのパッケージを紹介する。なおほとんどのものは CRAN(https://cran.r-project.org/) から入手可能である。

【TIMSAC】

TIMSAC(TIMe Series Analysis and Control program) は、赤池弘次元所長を中心として開発された時系列データの解析、予測、制御のための 総合的プログラムパッケージである。オリジナル TIMSAC(TIMSAC-72) は 1972 年に発表され、その後、TIMSAC シリーズとして TIMSAC-74,TIMSAC-78,TIMSAC-84 が Computer Science Monograph に発表された。工業プロセスの最適制御、経済変動の分析等広い分野で現在でも実際に利用されている。TIMSAC の特徴としては、情報量規準の考え方を用いた時系列解析プログラムであることが挙げられる。TIMSAC-72 では FPE(Final Prediction Error)、TIMSAC-74 以降では AIC(Akaike Information Criterion)、TIMSAC-78 のベイズ型モデルでは ABIC(Akaike Bayesian Information Criterion) も用いられている。

本パッケージは、FORTRAN で書かれているオリジナルプログラムの計算処理機能の多くをライブラリ化し、R関数を通して入出力を行い、必要であればその解析結果等をRでグラフィック表示することにより時系列データ解析を容易にしたものである。古典的な時系列解析パッケージとして今でも実際のデータ解析や時系列解析の教育に使われている。なお、過去のバージョンでこのパッケージに含まれていた七つの関数 armaimp(), lsar2(), ngsmth(), tsmooth(), tvvar(), tvar(), tvspc() は、最近のバージョンでは削除した。現在、これらの関数は次に示すRパッケージ TSSS の中で公開している。

【TSSS】

Rパッケージ TSSS は、北川源四郎元所長による書籍「FORTRAN 77 時系列解析プログラミング」(岩波コンピュータサイエンスシリーズ、岩波書店、1993) に掲載されていたプログラムを基に作成された時系列データ解析のための関数群である。「FORTRAN77 時系列解析プログラミング」では、代表的な時系列のモデリングに必要な最小二乗法、最尤法、カルマンフィルタによる推定の方法、情報量規準AICを用いたモデルの評価・選択の方法およびそれらを実現するプログラム等が紹介されている。現在は改訂版として、北川源四郎著「時系列解析入門」(岩波書店、2005) が出版されており、そこではFORTRANのソースコードは除かれているが、モデルや解析法について前書と同様に解説されている。TSSS は前書のデータをデータセットとして組み込んでおり、関数のドキュメントにおける例題の一部ではこれらのデータセットを用いている。

なお、時変係数ARモデルの時変分散と時変AR係数を推定する関数 (tvvar, tvar) については、OpenMP を使った拡張パッケージ tvvarOMP を利用して並列処理も可能にした。

【CATDAP】

CATDAP (CATegorical Data Analysis Program) は、坂元慶行名誉教授を中心に開発された最適な分割表（クロス表）の探索のためのプログラムである。最適な説明変数の選択には、AIC (Akaike Information Criterion) が使われている。CATDAP には、CATDAP-01 と CATDAP-02 の２つのプログラムがある。 CATDAP-01 は、カテゴリカルな（質的な）データに対し二次元分割表の比較を行い各変数の間の関係の深さを検出するプログラムである。全ての変数がカテゴリカルであることが前提となる。CATDAP-02 は、ひとつの着目した項目（目的変数）を固定し、他の項目（説明変数）の組合せで多次元分割表を作り、目的変数の分布の違いを最も適切に説明する説明変数の組合せを探索する。このプログラムは量的な項目も適当に区分してカテゴリカルなデータに帰着することにより、変数が量的か質的かにかかわらず適用できる。

Rパッケージ catdap は、FORTRAN で書かれた CATDAP の計算処理機能をライブラリ化することにより、R からこれらを関数として利用できるようにした。このパッケージには catdap1 と catdap2 の二つの関数があり、それぞれ CATDAP-01 と CATDAP-02 と同様の解析と結果出力をする。Rはオリジナルデータの変換も容易で分割表のモザイクプロット表示も可能なので、Rのパッケージとしたことで CATDAP での解析も効率良く行えるようになった。

最近、主として石黒真木夫名誉教授によりいくつかの拡張が行われた。現在の catdap2 では

- Base AIC (＝説明変数なしモデルの AIC) の利用ができる
- 連続値目的変数に適用できる
- 目的変数, 説明変数に欠測が含まれるデータに適用できる

ようになっており、オリジナルの CATDAP-02 が機能強化されている。これによりこのプログラムであてはめられたモデルと、目的変数に正規分布を仮定する回帰モデルあるいはロジスティック回帰モデルとの AIC の比較が可能になり、石黒氏によれば、”CAT” が ”TIGER” となった TIGERDAP (The Integrated GenERal Data Analysis Program) とでも称すべきバージョンとなっている。

なお、関数 catdap2 の機能をさらに使い易くするため、パッケージ R commander (Rcmdr) を使ったメニューインタフェースも利用可能である。

【NScluster】

Rパッケージ NScluster は、ネイマン･スコット型空間クラスターモデルのシミュレーションとパラメータ推定のための関数群である。 これらの関数は U.Tanaka, Y. Ogata and K. Katsura, Simulation and estimation of the Neyman-Scott type spatial cluster models (Computer Science Monographs, No.34, 1-44, The Institute of Statistical Mathematics, 2008) の FORTRAN プログラムをもとに開発された。ここで利用できるモデルは、トーマスモデルとその拡張モデル（タイプA、タイプB、タイプC) および逆べき乗型モデルの5種類である。

パラメータ推定のためにシンプレックス法を用いているが、モデルによってはかなりの計算時間がかかる。そのため、この時間のかかる計算処理の部分を OpenMP を使って並列化している。OpenMP が利用可能な環境であれば、環境変数にスレッド数を設定して実行時間の短縮を図ることができる。

【Rhpc】

Rhpc は snow の流れをくむRの並列化のパッケージであるが、その特徴は

- 並列化のためには MPI ライブラリを用いるが，2GB以上のデータ処理に対応している
- 多くの部分をCでプログラムして実行速度を上げている
- RからMPI外部プログラムを利用し易くなっている

などである。

現在も改善中であり、最近の改善点は、Serialize の高速化、Windows版における MS-MPI v8.1 以降の MPI_Comm_spawn のサポートによる修正、などである。

【Rmpenv】

Rmpenv は任意精度による実数と複素数の四則計算および基本的な数学関数、さらに行列積や逆行列を求める関数などを実現するパッケージである。現在機能拡張中であり、まだ CRAN には公開していない。