はじめに

何について書かれた本か？

この本は **Stan**（スタン）というソフトウェアとそのR用のパッケージである **RStan**（アールスタン）を使って，統計モデリングを習得する本である．

統計モデリングとは確率分布を使った数理モデルをデータにあてはめて現象の理解と予測を促す営みだ．かつては「データが少ないために簡単なモデルしか構築できない」「計算に時間がかかる」「計算の数式と実装が難解である」といった理由で実用的なデータ解析の手法とはならなかった．しかし今は違う．大規模のデータが入手しやすくなり，コンピュータの計算速度は向上し，さらに統計モデリングに特化したプログラミング言語の開発も進んでいる．このため現在では，統計モデリングは極めて有効なデータ解析手法となっている．

本書では統計モデリングを行うにあたってStanという新しいプログラミング言語を使う．Stanは優れたアルゴリズムを搭載し，開発も急速に進んでおり，R用のパッケージであるRStanが並行して公開されているためRから手軽に利用できる．Stanでは重回帰やロジスティック回帰など基本的なモデルはもちろんのこと，階層ベイズモデル，状態空間モデル，トピックモデルなどの高度なモデルもわずか30行程度のコードで書くことができる．さらに解析者の課題にあわせた拡張が簡単に可能だ．本書でStanを介して身につけた考え方は，Stanの文法が少々変わっても，そしていつかStanに代わる新しい統計モデリングフレームワークが出てきたとしても大いに役に立つと確信している．

前提とする知識

統計モデリングを厳密に実行しようとすれば，様々な知識と技術が要求される．例えば，統計学（特に確率分布）の知識，現象のしくみを考えて数式に置き換えるモデル化の能力，そしてその数式をコードに落とし込むプログラミング能力が必要となるだろう．またモデル化については経験による部分も大きい．本書の目標はこれらの知識や能力を伸ばすことにある．とはいえ紙幅の都合上，本書では，読者が以下の基礎的な知識と技能を有していることを前提としている．

- 確率と統計の基本的な知識：確率，確率分布，確率密度関数，条件付き確率，同時分布，周辺分布，相関係数，回帰分析に関する基礎知識．検定や区間推定の知識はなくてもよい．もしも確率や分布，回帰分析に関する知識が十分でないのであれば，[19] を一読することを勧め

たい．

- プログラミング能力：R 言語の初歩，すなわちデータ加工や作図が一通りできること．本書ではデータフレームの操作や，`plot()` 関数（あるいは ggplot2 の描画関数）による作図を頻繁に行う．R の操作に自信のない読者は [8] を一読するとよいだろう．
- 線形代数のごく基本的な知識：ベクトルや行列の演算に関して大学の教養レベルの知識があること．

なお統計モデリングを実際に行った経験は問わないが，モデルを構築し検証するには手間と時間が必要となるので，こうした作業を楽しもうという心構えや意気込みを読者には期待したい．

本書のあらすじ

本書は統計モデリングのハンドブックではなく，統計モデリングのチュートリアルを目指した．すなわち，最初の章から順番に通読すると Stan による統計モデリングの知識と技術が学べるように構成している．本書の構造は図 1 のようになっている．

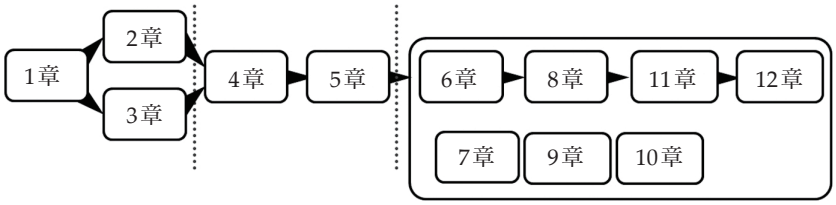

図 1　各章のつながり

大きく分けて 3 つの部からなる．1〜3 章は導入編であり，統計モデリングやベイズ推定などの理論的な説明を主にしている．4〜5 章は Stan 入門編であり，Stan そのものを取り上げ，単回帰や重回帰などの基本的な回帰分析を通してゼロから Stan と RStan の使い方を丁寧に説明する．6 章以降は発展編であり，Stan を使いこなす上で必須となる題材を取り上げた．6・8・11・12 章が本流であり，モデルのレパートリーを増やす章になっている．対して，7・9・10 章はモデルを改善する章となっている．特に発展編では読者は Stan の強力さを実感するだろう．

以下，より詳しく各章で扱っている内容を説明する．

1 章では統計モデリングと Stan の概要および特徴を簡単に説明している．

2・3 章ではそれ以降の章を読み進めるのに必要なベイズ推定と MCMC の用語，統計モデリングの手順を簡潔にまとめている．

4 章では Stan のインストール方法や基本的な文法を説明する．そのあとで，単回帰の例題を通して，実際に RStan を用いて R から Stan を実行して MCMC サンプルを得て，ベイズ信頼区間やベイズ予測区間を計算する．推定結果の見方と MCMC の設定の変更方法についても詳しく説明する．

5 章では 3 章の統計モデリングの手順に従って，重回帰やロジスティック回帰などの例題を扱う．また，図によるモデルのチェック方法についても説明する．

6 章では Stan そのものの説明からは少し外れるが，統計モデリングにおいて基本的な部品となる確率分布を紹介する．特に使われることが多いと思われる分布を取り上げた．

7 章では回帰分析を現実の問題に適用する上で悩みやすいポイントをまとめた．

8章では階層モデル（マルチレベルモデル）を導入する．これはグループ差や個人差を考慮した手法であり，今後利用される機会が増えるであろうモデルの一つである．

9章ではモデルをシンプルに記述するために利用できるデータ型の説明，高速化の方法，パラメータの制約，様々なエラーへの対処について説明する．

10章ではMCMCが収束しない場合の対策，特に弱情報事前分布について説明する．

11章では離散値をとるパラメータを扱う方法を説明する．現在のStanにはこうしたパラメータを処理できない弱点があるが，その回避方法を解説する．

12章では時系列データや空間構造があるモデルについて説明する．

本書で使用するソースコード

各章で使用するソースコードとデータについては，作図のコードも含めてほぼすべて本書のGitHubリポジトリ上(https://github.com/MatsuuraKentaro/RStanBook)で公開している．また，一部の章には実際に手を動かして理解を補うために章末に練習問題を付けた．これらの練習問題の解答も上記のGitHubリポジトリに置いたので，適宜参照してほしい．なお，本書では基本的に以下の3つの番号を一致させている．

- モデルを表す数式（モデル式`X-Y`）
- Stanコード（`modelX-Y.stan`）
- 実行するRコード（`run-modelX-Y.R`）

本書を執筆するにあたって使用した計算機環境は，Windows 7 (64 bit)，R 3.3.1，Rtools34，Stan 2.11である．