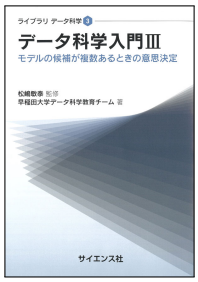
書評「データ科学入門Ⅲ」
モデルの候補が複数あるときの意思決定



松嶋敏泰 監修
早稲田大学データ科学教育チーム 著
サイエンス社（2024 年）
A5 判，176 ページ，
定価 2,090 円
ISBN：978-4-7819-1598-2

データ科学を支える専門分野として統計学と（機械学習を中心とした）情報科学がある．それぞれ，異なるコミュニティがあり独立して発展しつつも，互いに大きな影響を与えている．

データ科学において，「モデル」の定義は重要である．単回帰の「モデル」y=a+bx において，パラメータ a と b を求めることは単純な計算である．しかし，多項式の「モデル」の集合を考えるとき，適切な次数を決定することは簡単な問題ではない．古典的な統計学では，変数選択法を含む「アルゴリズム」が提案されてきた．特に，AIC は，標準的な情報量規準として広く使われている．また，BIC も情報量基準として広く使われている．ここで，「基準」と「規準」と書き分けているのは，小西 (2019) [1] が参考になる．

本書では，漢字の使い分けはしていないが，「構造推定の問題」として BIC，「予測の問題」として AIC を説明しており，これにより，AIC と BIC が，単に第 2 項が 2 ×（自由パラメータ数）と logn ×（自由パラメータ数）のどちらであるかという形式的な違いではないことが分かる．これらの体系化により，必要に応じて，WBIC などを勉強するための道標となる．統計学を含めた初学者にとってレベルが高いが，広い意味での「回帰」をモデルの立場から体系的に示している．

第 1 章は，本書の立場の外観を与えている．この章に目を通すことで，著者の立場が分かると思う．ただし，多くの読者にとって，この章だけを独立して理解することは困難である．その場合，軽く目を通して，第 2 章以降に進む方が良いと思われる．この章は，全体を読んでから，または，より学習が進んでから，時間をかけてじっくりと味わうことをお勧めする．

第 2 章は，回帰における基礎事項を分かり易く扱っている．多項式回帰や重回帰分析から始めて，リッジ回帰，lasso 回帰など縮小推定まで紹介している．かなり詳細な内容を含んでいる．

第 3 章は，「モデル」の扱いについての中心的な章である．分野や研究者によって「モデル」の定義が異なるので，著者の「モデル」の定義を正確に把握することが望まれる．「モデルが既知」と「モデルが未知」の違いを丁寧に解説したうえで，後者における扱いの基礎を扱っている．ところどころ，ベイズアプローチにもふれている．BIC，AIC の違いを理解されたい．変数増加法・変数減少法などのモデルの探索アルゴリズムも紹介している．

第 4 章では，（本書における）同質性を仮定した場合の予測について扱っており，クロスバリデーションのいくつかの手法，データ科学において基本的な事項である正則化の枠組みでリッジ回帰，lasso 回帰などを簡潔に紹介している．

第 5 章では，ニューラルネットワークの基本を説明したのち，CNN やディープニューラルネットワークまで丁寧に扱っている．深層学習を応用する際，単にパッケージを利用するだけのエンドユーザ以上の理解が可能になる．

付録も，深層学習などを学ぶための数理として有用である．

参考文献では，数理的な立場から統計学を学ぶための良書がそろっている．

この分野の進展は急激である．データ科学は，まさに変容・進化しており，新しい手法や理論が次々と登場している．本書でも少しだけ紹介している生成 AI が，社会において大きな影響を与えている．また深層学習などで登場する Double descent（二重降下）は，表面的には，Bias-Variance Trade-off と反することである．このような時代におけるデータ科学の学習においては，統計学・情報科学を含めた基本的な概念の学習が求められている．そのような意味で，統計学と機械学習の架け橋となる本書は，名著だと考える．

Received: 14 November 2024.

参考文献

[1] 小西貞則 (2019)：情報量規準 AIC の統計科学に果たしてきた役割「5. 赤池ベイズ情報量規準 ABIC とベイズ型モデル評価基準 BIC」，統計数理，第 67 巻第 2 号，pp. 193-214 https://www.ism.ac.jp/editsec/toukei/pdf/67-2-193.pdf

（統計数理研究所　水田 正弘）