

まえがき

21世紀はデータの世紀と言われている。我が国の科学技術基本計画では、Society5.0という名のもとに第4次産業革命を展開することが謳われている。これはイギリスの蒸気機関の発明による第1次産業革命、アメリカのモーター・ベルトコンベヤー発明による第2次産業革命、日本を中心としたエレクトロニクスによる第3次産業革命に続く第4次産業革命と位置付けられている。そこではIoT技術の高度化と普及を背景として、通信ネットワークやセンサーなどのデータ収集機器を通じたビッグデータを活用し、従来の経験と勘による意志決定を脱して、データ分析による科学的な意思決定を目指す社会ということができ、データの分析がその中心に位置づけられる。

これらを背景として、データサイエンスやAI（人工知能）・数理教育を、文系理系を問わずに入学初年度から実施する取組みが多くで大学で行われている。統計教育も大きな分岐点を迎えている。経済経営分野の統計教育は歴史と伝統のある裾の広い分野であり、計量経済学を核とした経済経営データ分析の体系がある。機械学習やAIの進展はデータ分析の視点と範囲を広げ、経済経営分野でもこれらを積極的に取り入れることが必要な時代になっている。

このねらいのもと、本書は統計学を基礎としながら新しいデータ分析の手法を体得することを目的として書かれた。本書では主にビジネスデータで活用されてきた多変量解析と呼ばれる伝統的手法に加えて、ベイジアンネットワークやLassoなどの代表的な機械学習や初等レベルのAI手法を用いて、経済経営データを分析しながら学ぶ。近年では学生が自身のノートPCやタブレットを教室に持ち込んで講義を受け、その場で実習を行うBYOD（bring your own device）環境を前提とする大学も少なくない。本書は、これへの対応も意識して、各章で説明される内容や例題を自身のデバイスで追体験できるようにした。

本書は、確率統計の基礎知識を前提とし、統計学入門を履修した大学2年

vi | まえがき

次以降の学部生，大学院生および経済経営のデータ分析に関心を寄せる実務家を対象にしている。本書の構成は，回帰分析の基礎から始まり，その後のトピックはほぼ独立に学ぶことができる。各章の例題で使われたデータと分析の R コードは，<https://www.kyoritsu-pub.co.jp/bookdetail/9784320125193> からダウンロードして利用できる。R はフリーにダウンロードできる統計分析ツールであり，すでに多くの解説書がある。大学によっては入学初年度からのデータサイエンス・AI 教育で使い方を学べることも多い。各章の理解に際しては，自身でデータ分析を実践し，データから情報を抽出することがいかに容易で面白いかをぜひ体得してほしい。

著者一同

目 次

第 1 章 線形式で予測する：回帰モデル	1
1.1 単回帰モデル	1
1.1.1 単回帰モデルの定義	1
1.1.2 最小二乗法	4
1.2 重回帰モデル	6
1.2.1 重回帰モデルの定義	6
1.2.2 最小二乗法	7
1.2.3 重回帰モデルの解釈	9
1.2.4 ダミー変数	12
1.2.5 予測値と残差	15
1.2.6 決定係数	17
1.3 最小二乗推定量の性質	21
1.3.1 最小二乗推定量の期待値	21
1.3.2 除外変数バイアス	24
1.3.3 最小二乗推定量の分散	25
1.3.4 最小二乗推定量の有効性：Gauss-Markov の定理	26
第 2 章 変量間の関係を調べる：回帰モデルの統計的推測	29
2.1 最小二乗推定量の標本分布	29
2.2 t 検定	31
2.2.1 t 分布	31
2.2.2 片側対立仮説	33
2.2.3 両側対立仮説	38
2.2.4 0 以外の帰無仮説の検定	40
2.2.5 p 値	42

viii | 目 次

2.2.6	信頼区間	44
2.3	F 検定	45
2.3.1	複数パラメーターの同時検定	46
第3章	モデルの複雑さをコントロールする：正則化	53
3.1	行列表記	53
3.1.1	内積とノルム	53
3.1.2	行列の復習	54
3.1.3	回帰モデルの行列表記	55
3.1.4	行列表記による最小二乗（OLS）推定量の導出	57
3.2	多重共線性	58
3.2.1	完全な多重共線性	58
3.2.2	近似的な多重共線性	59
3.3	正則化	61
3.3.1	最小二乗法の限界	61
3.3.2	正則化	62
3.3.3	正則化推定量の別の定義	63
3.3.4	変数選択と Lasso	65
3.3.5	正則化係数の選択	66
3.4	バイアスと分散	67
3.4.1	平均二乗誤差の分解	68
3.4.2	正則化推定量の性質	69
第4章	高次元回帰モデルを効率的に推定する：Lasso	71
4.1	Lasso 推定量の性質	71
4.1.1	スパース回帰モデル	71
4.1.2	誤差上限	72
4.1.3	変数選択	73
4.2	adaptive Lasso	74
4.2.1	オラクル性	75

4.3	実証分析	76
4.3.1	R コード	76
4.3.2	データ	77
4.3.3	推定結果	79
第 5 章	高次元における統計的推測：多重検定	85
5.1	debiased Lasso 推定量	85
5.1.1	漸近正規性	86
5.1.2	標準誤差の推定	88
5.1.3	R コード	88
5.2	多重検定	89
5.2.1	検定論の復習	89
5.2.2	ファミリーワイズエラー率	91
5.2.3	Bonferroni 法	92
5.2.4	FDR と BH 法	93
5.2.5	実証分析	95
第 6 章	統計手法が正しく機能するか調べる： モンテカルロ実験	97
6.1	期待値の検定のモンテカルロ実験	98
6.1.1	母集団が正規分布の場合	98
6.1.2	母集団の分布が不明な場合 (標本数がある程度大きい時)	103
6.1.3	母集団の分布が不明な場合 (標本数が小さい時)	107
6.2	回帰係数の検定のモンテカルロ実験	111
6.2.1	家賃データでの分析	112
6.2.2	モンテカルロ実験の設定	115
6.2.3	モンテカルロ実験の結果	117
6.3	現代の統計学におけるモンテカルロ実験	119

x | 目 次

第7章 コンピューターの力で難題を解決する：	
ブートストラップ	122
7.1 期待値の検定のブートストラップ	123
7.1.1 家賃データでの分析	123
7.1.2 モンテカルロ実験	126
7.1.3 補足	127
7.2 回帰係数の検定のブートストラップ	129
7.2.1 家賃データでの分析	129
7.2.2 モンテカルロ実験	131
7.2.3 補足	132
7.3 現代の統計学におけるブートストラップ	133
第8章 データを可視化する：	
主成分分析，因子分析，多次元尺度構成法	136
8.1 主成分分析	137
8.1.1 主成分分析の考え方	137
8.1.2 分析事例：ホテル利用者の評価による 市場ポジションの可視化	138
8.2 因子分析	141
8.2.1 因子分析の考え方	141
8.2.2 因子分析のモデル	142
8.2.3 分析事例：歯磨きブランドの属性因子の可視化	143
8.3 多次元尺度構成法	146
8.3.1 多次元尺度構成法の考え方	146
8.3.2 分析事例：家計消費から見た都道府県の 類似性の可視化	148
8.4 補論：主成分分析の数理	148

第9章 集団を分類する：	
クラスター分析，ナイーブ・ベイズ分類，決定木	151
9.1 クラスター分析	151
9.1.1 クラスター分析の考え方	152
9.1.2 分析事例：生活指標による国の分類	153
9.2 ナイーブ・ベイズ分類	155
9.2.1 ベイズの定理と分類問題	155
9.2.2 ゼロ頻度問題とベイズ確率モデル	156
9.2.3 分析事例：広告バナー・クリックに関する ユーザー分類	157
9.3 決定木	158
9.3.1 決定木の考え方	158
9.3.2 分析事例：旅行プラン採用集団の分類とルールの抽出	161
9.4 補論：共役事前分布とベイズ確率モデル	164
第10章 判別して要因を探る：	
ロジスティック回帰，判別分析	166
10.1 ロジスティック回帰	166
10.1.1 ロジスティック回帰の考え方	166
10.1.2 分析事例：顧客満足の判別と要因	167
10.2 判別分析	171
10.2.1 事後確率最大化と線形判別関数	171
10.2.2 分析事例：学生のオンライン授業選択の判別と要因	172
第11章 原因を推定する：ベイジアンネットワーク	176
11.1 ベイジアンネットワークとは	176
11.1.1 現象から原因を探る：ベイズの定理	176
11.1.2 ベイジアンネットワーク	178
11.1.3 ベイジアンネットワークの特徴	182

xii | 目 次

11.2 ベイジアンネットワークの使い方 184
 11.2.1 ネットワークの構造を決める 184
 11.2.2 原因が生じる確率の推論 186
 11.3 分析事例：ダイレクトマーケティング分析 186

第 12 章 文書から話題を見つける：トピックモデル 190

12.1 文書データの分析 190
 12.1.1 構造化データと非構造化データ 190
 12.1.2 文書データ分析の難しさ 191
 12.2 トピックモデル 192
 12.3 潜在ディリクレ配分法 193
 12.3.1 文書生成の統計的生成モデル 193
 12.3.2 LDA の数理モデル 194
 12.3.3 トピックモデル分析のための前処理 196
 12.3.4 分析事例：LDA によるレビュー分析 197
 12.4 STM：構造トピックモデル 198
 12.4.1 STM の数理モデル 198
 12.4.2 分析事例：STM によるレビュー分析 200

第 13 章 好みを見つける：推薦システム 205

13.1 推薦システム 205
 13.1.1 推薦システムとは 205
 13.1.2 推薦システムとマーケティング 206
 13.1.3 推薦システムの種類 208
 13.2 協調フィルタリング 209
 13.2.1 問題の定式化 209
 13.2.2 ユーザーベース協調フィルタリング 210
 13.2.3 Matrix Factorization 212
 13.3 分析事例：映画の推薦 213
 13.4 より進んだトピック 216

参考文献	219
略 解	223
索 引	227