

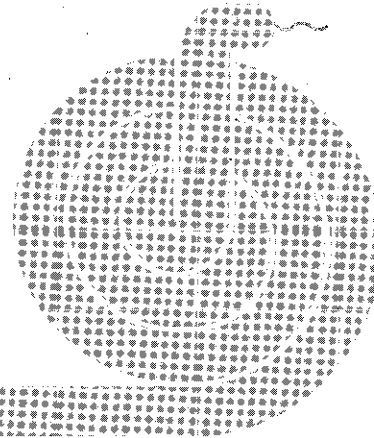
いまさら聞けない

Pythonでデータ分析

多変量解析, ベイズ統計分析
(PyStan, PyMC)

OKAMOTO YASUHARU

岡本安晴



丸善出版

はじめに

本書は、データ分析をPythonで行うときの基礎的事項の解説を試みるものである。Pythonは、プログラミング言語としてはやさしい使い方もできるが、いろいろなライブラリ・パッケージが用意されているので、様々な分野で用いられている汎用性の高い言語でもある。

読者として、初めてPythonでデータ分析を行ってみようと思う人を想定している。Pythonは全く初めてという人も本書でPythonが使えるように工夫したが、言語の文法の詳しい解説は他書に譲った。データ分析に関わる基礎的事項はできるだけ扱うようにして、本書を辞書的に参照して長く利用されるようにした。

まず、データおよびその分析結果の視覚化としてPythonのライブラリmatplotlibを用いたグラフ描画について説明するが、これは本書全体にわたる分析の視覚化において用いられる。データ分析法としては、基本となる標準的な多変量解析と現在注目を集めている確率モデルによるベイズ分析を取り上げる。多変量解析においては行列演算を用いるのが現在の標準的方法であるが、ライブラリnumpyを用いた行列演算について説明する。ベイズ分析については、StanのPython用であるPyStan、およびPython専用のPyMCを取り上げる。いずれも、基礎的使い方の説明を心がけた。

各章の内容は以下のとおりである。

第1章 基本統計量の計算

平均、分散など基本統計量を取り上げて、Pythonのコード例を示した。Pythonプログラミングの雰囲気を示すとともに、numpyなどのライブラリの利用例も示した。

第2章 グラフ描画——データの可視化——

情報は視覚化されるとわかりやすい。データおよびその分析結果の視覚化はグラフという形で表示される。Pythonでのグラフ描画のライブラリとしてmatplotlibを取り上げる。棒グラフ、ヒストグラム、折れ線グラフを例に挙げて、その使いやすさを示す。

第3章 ファイル入出力

簡単なデータは、Python スクリプト中に書き込むことができるが、実用的にはファイルから、あるいは適当なストリームからの入力になる。本章において、ファイル入出力の方法として、テキストファイル、CSV (Comma-Separated Values)形式ファイル、バイナリファイルを取り上げる。

第4章 行列演算とPython スクリプト

多変量解析の基礎である行列演算を取り上げる。Python における行列演算のライブラリとして numpy を取り上げ、行列演算が簡単に行えることを説明する。

第5章 単回帰分析

変数間の関係を1次式で表すのは、統計分析の1つの出発点である。その最も簡単なモデルが2変数の関係を1次式で表す単回帰モデルである。この単回帰モデルを行列演算で表し、その解が行列で簡単に表せることを説明するとともに、numpy の ndarray を用いたスクリプト例を示す。

第6章 重回帰分析

重回帰分析では、複数の独立変数の影響が1次式で表される。行列を用いると、形式的には単回帰モデルと同様に扱える。しかし、複数の独立変数を用いることにより、単回帰モデルでは表されない関係を扱うことができるが、これは演習課題とした。

第7章 主成分分析

多くの変数から構成されるデータは、その主な情報がいくつかの成分で表されることが多い。この成分を変数の1次式で求める方法として主成分分析があるが、本書ではこれを正射影の観点から説明した。主成分分析も行列を用いると簡単に表すことができる。

第8章 数量化

調査あるいは質問紙データは、カテゴリ変数がよく用いられる。カテゴリ変数を数量化すると、他の数量変数とともに標準的な多変量解析を適用することができる。数量化の計算は、行列を使えば簡単である。数量化のためのPython スクリプト例を示す。

第9章 確率計算とPython スクリプト

ベイズ分析は、確率モデルの強力な方法として注目を集めている。確率計算のPython スクリプト例とともに、ベイズ分析の基本モデルの簡明さを示す。

第10章 PyStan による2項分布分析——Stan 入門——

ベイズ分析は、事後分布をシミュレーションで求める方法が開発されて実用性と有効性が飛躍的に高まった。Stan は、シミュレーションで求めるライブラリの1つであり、Python 用のStan がPyStan である。基礎的な使い方は簡単であり、簡単な確率モデルである2項分布を使って説明する。

第11章 PyStan による単回帰モデル分析

統計分析の基本モデルとして単回帰モデルを取り上げ、PyStan によるベイズ分析について説明する。

第12章 PyStan によるポアソン回帰モデル分析

回帰モデルの一般化の1つとして、ポアソン回帰モデルを取り上げる。また、分散分析のように要因がカテゴリ変数のときは、ダミー変数を用いると回帰モデルが設定できる。クロス表の分析を、ポアソン回帰モデルにおいてダミー変数を設定して行う。

第13章 PyMC による2項分布分析——PyMC 入門——

PyMC によっても、ベイズ分析における事後分布をシミュレーションで求めることができる。基本的方法是簡単であり、簡単な確率分布として2項分布モデルを取り上げて説明する。

第14章 PyMC による単回帰モデル分析

統計分析の基本モデルの1つとしての単回帰モデルを取り上げ、PyMC による分析について説明する。

第15章 PyMC によるポアソン回帰モデル分析

回帰モデルの一般化の1つとして、ポアソン回帰モデルを取り上げる。また、分散分析のように要因がカテゴリ変数のときは、ダミー変数を用いると回帰モデルが設定できる。クロス表の分析を、ポアソン回帰モデルにおいてダミー変数を設定して行う。

本書の企画段階から1次原稿まで、いろいろな方々から御意見、御提案、励ましをいただいた。ここにすべての方のお名前を挙げることはできないが、謝意を表する次第である。特に大阪大学の足立浩平教授および狩野裕教授の研究室の方々からコメントやアドバイスなどをいただいたことは記しておきたい。また、加藤直哉氏と加藤仁美氏には、初校前の原稿について励みとなるコメントをいただいたことを記しておきたい。もちろん、本書に残る誤りなどの問題点は、最終的に著者の責任であることは言うまでもない。

丸善出版株式会社企画・編集部第三部長の小西孝幸氏には、筆者の最初の企画提案の段階から有益な助言をいただいた。本書の現在の内容は、氏の援助に基づくところが大きい。著者は当初、データ分析・統計分析の分野に進もうという学生でプログラミングが初めてという人を対象とした Python の入門書を考えていた。小西氏から Python 関連書籍の情報などが提供され、著者の研究および授業におけるプログラミングの経験をよりよく反映した本書を上梓することができた。ここに記して謝意としたい。

2018 年 夏

自宅にて
岡本安晴

目 次

第 1 部 Python データ分析入門	1
序 章 Python の準備と使い方	2
序.1 準備(インストール)	2
序.2 使い方	4
第 1 章 基本統計量の計算	8
第 2 章 グラフ描画——データの可視化——	14
2.1 棒グラフ	14
2.2 ヒストグラム	16
2.3 散布図	18
2.4 折れ線グラフ	23
2.5 ラインスタイルの設定	25
コラム 2.C.1 疑似相関	26
演習課題	27
第 3 章 ファイル入出力	28
3.1 テキストファイル入出力	28
3.2 CSV 形式	34
3.3 バイナリファイル入出力	36
第 2 部 多変量解析	37
第 4 章 行列演算と Python スクリプト	38
4.1 行列の表現	38
コラム 4.C.1 リストの初期化	43
4.2 四則演算	44
4.3 トレース・階数・ノルム	52
4.4 固有値と固有ベクトル	61
4.5 特異値と特異ベクトル	67
4.6 行列式	72

第5章 単回帰分析	75
5.1 モデル	75
5.2 Python スクリプト	79
演習課題	88
第6章 重回帰分析	90
6.1 モデル	90
コラム 6.C.1 ダミー変数のコーディング	92
6.2 Python スクリプト	94
6.3 標準化回帰係数	105
演習課題	114
コラム 6.C.2 ダミー変数の独立性	114
第7章 主成分分析	115
7.1 モデル	115
7.2 Python スクリプト	118
第8章 数量化	130
8.1 モデル	130
8.2 Python スクリプト	133
第3部 ベイズ分析	147
第9章 確率計算とPython スクリプト	148
9.1 離散確率分布	148
9.2 連続確率分布	153
9.3 乱数・確率分布・シミュレーション	159
9.4 ベイズ分析	169
コラム 9.C.1 事後分布の記述	176
第10章 PyStan による2項分布分析——Stan 入門——	178
第11章 PyStan による単回帰モデル分析	189
演習課題	200
第12章 PyStan によるポアソン回帰モデル分析	202
コラム 12.C.1 一般化線形モデルとリンク関数	208
第13章 PyMC による2項分布分析——PyMC 入門——	211
第14章 PyMC による単回帰モデル分析	217

演習問題	227
第15章 PyMC によるポアソン回帰モデル分析	228

参考文献	237
索引	240