まえがき

この本は、初学者のために書かれた時系列分析の入門書です。

パラメタ推定など技術的な面は一部省略されている代わりに、時系列データの構造や、様々な時系列モデルの特徴、ライブラリを用いた分析の手順や、結果の解釈の仕方に紙数を割きました。

Box-Jenkins 法と状態空間モデルを共に学べるのも大きな特徴です。初めて学ぶ方でも、古典的な内容からステップアップするため無理なく学べるはずです。

ビッグデータという言葉が代表するように、とても多くのデータが分析に使われるようになりました。しかし、どれほど多くのデータを持っていても、決して手に入らないデータがあります。

それは未来のデータです。

100 テラバイトの売り上げデータを持っていても、高頻度証券データを取得したとしても、1 ミリ秒先の未来のデータは、私たちの手元にありません。

まだ手に入っていない未来のデータについて言及することが予測であり、時系列分析は、予測を行う強力なツールです。そして、状態空間モデルは、現代の時系列分析の事実上の標準ともいえるフレームワークです。

この本では「R」や「Stan」といった無料のソフトウェアを用いて、時系列データを効率的に分析する方法も説明します。

統計学は、無から有を生み出す錬金術ではありません。

常に予測ができるわけではなく、どうしても予測の出しようがないこともあります。また、時系列データ特有の問題として、素朴な回帰分析などを適用すると、誤った考察を導いてしまうこともあります。

時系列データを分析する際の注意点、逆に、時系列データだからこそ持っている情報を利用する方法。両者をバランスよく学び、時系列データを有効活用してください。

統計学は便利な道具です。統計学を教える書籍も便利な道具であるべきです。

本書が皆さんにとって、有用なツールとなることを願います。

本書の構成

この本の難易度

この本では、時系列分析のアイデアを伝えることに注力しました。統計学にそれほど詳しくなくても最後まで読めるように配慮されています。数式の量も、この分野の標準教科書と比べるとかなり抑えられているはずです。

しかし、時系列分析は統計学の応用編ともいえる分野です。推定や検定、回帰分析や最小二乗法といった言葉をある程度知っている、という方がこの本を読まれると良いでしょう。

この本の読み方

節のタイトルにアスタリスク（*）がついているものは、テクニカルな話題となるため初学者が読むのはやや難しい可能性があります。

アスタリスクの有無にかかわらず、少し難しいかなと思った個所は（数式も含めて）どんどん飛ばしていき、自分がわかる部分だけをかいつまんで読むというのも、良いやり方だと思います。特に数式は、必ずその解釈を日本語で書くようにしているので、たとえ飛ばしたとしてもある程度は理解できるはずです。

また、理論の説明をした後でソフトを使って実装するという説明の仕方で統一されています。難しいと感じた理論は軽く読み流したうえで、自分で実装しながら都度読み返すという進め方をとることもできます。

計算のためのソフトウェアはすべて無料で手に入れることができます。ソースコードは著者の Web サイト（https://logics-of-blue.com/）から無料でダウンロードできます。

この本の構成

本書は 6 部構成となっています。目次をかなり詳細に書いてあるため、自分が今どこにいるのか、次に何を学ぶのかがわかるようになっています。

「第 1 部　時系列分析の考え方」では、時系列分析とは何かという基本から説明をします。

特に時系列データの構造、時系列モデルという考え方を理解してください。後

ほど学ぶ分析手法のほぼすべてで必要とされる考え方です

「第 2 部　Box-Jenkins 法とその周辺」では ARIMA モデルと呼ばれる古典的な時系列モデルを中心とした分析の方法を説明します。

Box-Jenkins 法は比較的古い手法とはいえ現代でも十分に実用的です。

また、Box-Jenkins 法は時系列分析の基礎を学ぶ格好の教材ともいえます。定常性や和分過程といった時系列データ特有の考え方に加え、モデル選択や残差のチェック、予測精度の評価といった分析における一般的な流れを学ぶことができるからです。ここまでを読了できれば、時系列分析の基本的な用語や考え方が身についているはずです。

「第 3 部　時系列分析のその他のトピック」は、独立した 3 つの章で構成されています。

1 つは時系列データに対して回帰分析を適用した際の問題点「見せかけの回帰」について。

2 つ目は多変量時系列データの分析手法としての VAR モデルについて。

3 つ目は分散不均一なデータへの分析手法として ARCH・GARCH モデルを解説します。

「第 4 部　状態空間モデルとは何か」でこの本のメインテーマの 1 つである状態空間モデルの導入をします。

「第 5 部　状態空間モデルとカルマンフィルタ」では、カルマンフィルタを用いて、線形ガウス状態空間モデルを推定します。カルマンフィルタや散漫カルマンフィルタ、平滑化などの考え方と計算方法を解説します。

(散漫)カルマンフィルタと平滑化に関しては、ローカルレベルモデルと呼ばれる単純なモデルを例に挙げて、ライブラリを使わずに自分の手で実装します。基礎を学んだ後、現実に近い問題を、ライブラリを用いて分析していきます。

「第 6 部　状態空間モデルとベイズ推論」では、MCMC を用いて非ガウシアン・非線形のモデルを構築します。ベイズ推論の基礎から、一般化状態空間モデルの実装方法までを解説します。

第1部　時系列分析の考え方

まずは、時系列データの特徴と時系列分析の考え方を解説します。

時系列データがもつ特徴、そして時系列データだからこそ生じる問題とその解決策を学びます。

1章　時系列分析の基礎

時系列分析とは何をするものか、という基本を解説します。

基本的なデータ分析の枠組みと変わらない部分、そして時系列分析特有の考え方を学びます。

1－1　　推測統計学の考え方

推測統計学では「標本から母集団を推定する」という言い方をします。

日本人の平均身長が知りたかったとしても、日本人全員の身長を測定するのには無理がありますね。

なので、ちょっと少ないですが、100 人とか 200 人とかをサンプリングして標本とし、その標本から日本人全員という母集団について議論します。

サンプリングされた 100 人の平均身長が 160cm だったとしましょう。すると「日本人全員の平均身長も 160cm ではないか」と推定できます。

また標本の身長が、130cm～190cm と大きくばらつく、すなわち、データの分散が大きかったとします。そうしたら、推定された平均身長もある程度ばらつきそうです。

そんなばらつきを加味して 95%信頼区間などを求める、ということをした経験があるのではないでしょうか。

手持ちのデータ、すなわち標本から期待値や分散といった統計量を計算し、まだ手に入れていないデータについて言及する。

これがいわゆる推測統計学の考え方でした。

1－2　　時系列データとは

時系列分析は、文字通り時系列データを取り扱う分析手法のことです。

時系列データは、例えば、神戸市の毎日の気温の推移であったり、毎年観測されるアザラシの個体数の推移であったり、飲食店の毎月の売上金額の推移であったりします。日、あるいは月や年、時・分・秒など一定の間隔で取られた、一連

のデータを時系列データと呼びます。
一方、時系列データでないデータを、区別するためにトランザクションデータと呼びます。

この章では、主に日単位のデータを対象として議論を進めていきますが、特に断りがない限り、これは月単位データなどほかの単位のデータでも同じように考えることができます。
また、このことを明示的に示すため、この本では、「1時点前」といったように「○時点」という表現をしばしば使います。
「1時点前」ならば、時系列の単位によって「1日前」になったり「1月前」になったり「1年前」になったりします。

1－3　時系列データをどのように取り扱うべきか

時系列データには大きな特徴があります。
それは「一日のデータ」は、「一日に一回しか手に入らない」ということです。
例えば、2000年1月1日という日は、この世界に1つしかありません。かけがえのない一日である2000年1月1日の神戸市の気温というデータが、私たちの手元にあったとしましょう。これが標本です。

標本は手元にありますね。
では、標本から母集団を推定しましょう。
ここで、大きな問題にぶつかります。
この場合の母集団とは、いったいなにものでしょうか。

もしも、2000年1月1日という日が無数にあったならば、「無数に存在する2000年1月1日の神戸市の気温」が母集団となります。
仮に母平均を推定しようと思ったら「無数に存在する2000年1月1日」という日の気温の平均を、「手元にあるたった一つの2000年1月1日」から推定しなければなりません。
これが、時系列データの持つ難しさです。

1－4　母集団と確率分布・標本と確率変数

母集団という存在は大変に扱いづらいです。そもそもこの母集団が一体何者かを想像することすら難しい。

ですが、推測統計学は、強力な武器を私たちに残してくれました。それが確率分布と確率変数という考え方です。

確率変数とは、確率的に変化する値です。

サイコロの例を挙げましょう。

6 分の 1 の確率で、三の目が出てきますね。出てくる目が確率的に変わるので、これは確率変数です。

確率分布とは、データが出てくる確率の一覧です。

いかさまでないサイコロの場合は、以下のようになります。

確率変数｛1,　2,　3,　4,　5,　6｝
確率分布｛1/6，1/6，1/6，1/6，1/6，1/6｝

推測統計学では、以下のように考えます。

- 標本とは確率変数である。
- 母集団として、ある特定の確率分布を想定する

サイコロの場合は｛1/6，1/6，1/6，1/6，1/6，1/6｝こそが母集団の確率分布です。標本は、この母集団の確率分布に従って得られると考えます。

1－5　データ生成過程(DGP)の考え方

データ生成過程(Data Generation Process：DGP)とは、時間に従って変化する確率分布のことです。確率過程、あるいは単に過程とも呼ばれます。

神戸の気温は、何らかの確率分布に従って得られるのだと考えます。

2000年1月1日の気温の確率分布は以下のようだったとします。

気温(℃) {1,　2,　3,　4,　5,　6}
確率分布 {1/6, 1/6, 1/6, 1/6, 1/6, 1/6}

データ生成過程がわかっていれば、2000年1月1日の気温の期待値は簡単に計算できます。3.5℃ですね。
もちろん分散も計算できます。

次に、翌日2000年1月2日の気温も、やはり何らかの確率分布に従って得られていると考えます。
ただし、確率分布が若干変わります。

気温(℃) {1,　2,　3,　4,　5,　6}
確率分布 {1/4, 1/6, 1/6, 1/6, 1/6, 1/12}

寒くなる確率が増えました。

私たちの手元には、2000年1月1日の気温が一つだけ、2000年1月2日という日の気温もたった一つがあるのみです。
そのデータを「本来ならばあり得た」確率変数の一つの実現値だとみなします。次にサイコロを投げる機会がもしあれば、きっと異なる目が出るだろうと考えるのと同様に“もしも今日という日が複数あれば”次には異なる気温が得られたかもしれないと想定するわけです。
データ生成過程からデータが得られたと仮定して、たった一つしかないデータから理論的な期待値や分散を求めます。

1－6　DGPと時系列モデル

データ生成過程がわかっていれば、期待値や分散が計算できるだけでなく、未来を予測することもできるでしょう。
次に取り組むのは、データ生成過程をどのようにして推定するのかという問題

です。

何の情報もないままにデータ生成過程を推定するのは難しいですが、現実のデータには、多くの場合何らかの構造があると想定できます。

例えば気温の場合、冬になると気温が下がり、夏になると上がるだろうと予想されます。
また、昨日が寒ければ翌日も寒くなりそうです。言い換えると、昨日の気温と今日の気温が似ていると予想されます。
また、地球温暖化により、徐々に気温が上がっていくというトレンドがあるかもしれません。

こういったデータ生成過程の構造をモデル化します。データ生成過程の構造のことを時系列モデルと呼びます。
時系列分析の大きな目的の一つは、この時系列モデルをデータから推定することです。
時系列モデルが推定できればデータ生成過程がわかり、データの理論的な期待値や分散を計算することができます。季節の影響やトレンドの有無などを判断することもできるでしょう。
また、時系列分析の大きな目的の一つである将来予測を行うツールとしても、時系列モデルは大きな役割を果たします。

2章　時系列データの構造

時系列モデルを作成するためには、時系列データの構造を知ることが第一です。
この章では、典型的な時系列データが持つ構造を解説します。

2－1　自己相関とコレログラム

時系列データの特徴は、データに前後の関係があることです。

自己相関とは、過去と未来の相関をとったものです。
正の自己相関があれば「昨日の気温が高ければ今日も高い」ということになり、逆に負の自己相関があれば「昨日の気温が高ければ今日は低い」ということになります。
自己相関の様子がわかれば、モデルの特定に役立つだけでなく、自己相関という情報を使って未来を予測することもできるでしょう。単純な話、昨日の気温と正の自己相関という情報があるのならば「昨日の気温が高かったので今日も高いだろう」と予測を出せます。

また、どれくらい離れた時期と相関があるかということも重要です。
1時点前と相関があるのか、7時点前と相関があるのか、あるいは365時点前と相関があるのか。これも考えながらモデルを構築します。
何時点前と強い自己相関があるのかを調べるために、自己相関をグラフにすることがよくあります。このグラフをコレログラムと呼びます。

2－2　季節成分・周期成分

例えば、ひと月に一回だけデータをとった時系列データがあったとします。毎月の平均気温データなどを想像してください。
このとき12か月前のデータと強い正の相関があったとしましょう。
これはもちろん自己相関として片付けてしまうのも一つの手ですが、より明確に「毎年周期的に変動している」とみなしてモデルを作るほうがベターです。
夏ならば気温が高く、冬になると気温が下がるというのは容易に想像がつきま

すよね。
年単位の周期性は「季節成分」あるいは「季節性」と呼ばれます。

1 日単位のデータでは、曜日によっても周期的にデータが変化することがあります。例えばおもちゃ屋さんの毎日の売り上げは、おそらく平日よりも休日に高くなるでしょう。
こういった「データが持っているかもしれない周期成分」は積極的にモデルに組み込んでいきます。

同じ自己相関であっても「昨日と今日がよく似ている」という特徴と「毎週土曜日に売り上げが高くなる」という特徴は明確に分けてモデルを構築するほうが好ましいといえます。
このほうが、分析結果を意思決定に活用しやすくなるからです。
例えば毎週土曜日によく売れるというデータの構造がわかったならば、予測を出すまでもなく、金曜日に商品をたくさん仕入れるという行動をとることができます。また、平日の顧客数を増やすための施策を打つことが必要だという認識を共有することもできるでしょう。
単に予測モデルを構築するだけでなく「データの特徴をモデルで明確に表現する」ことができた方が、応用の幅が広がるということは覚えておいてください。

2－3　トレンド

例えば商品の売れ行きが好調で、毎月売り上げが右肩上がりで増えていったとします。このような状態を「正のトレンドがある」と呼ぶこともあります。
毎月 20 万円ずつ売り上げが増えるトレンドがあれば、「来月の売り上げ＝今月の売り上げ＋20 万円」で予測できますね。

2－4　外因性

近くでイベントがあったので、飲み物が多く売れた、といったように、外部の要因が影響を与えることもあります。これを外因性と呼びます。
データの自己相関だけでは表すことのできない振る舞いを説明することがで

きます。

2－5　ホワイトノイズ

ホワイトノイズは「未来を予測する情報がほとんど含まれていない、純粋な雑音」だと考えるとわかりよいです。

ホワイトノイズが満たす要件は「期待値が 0 であり、分散が一定であり、自己相関が 0 である」ということです。

平均 0、分散σ^2の正規分布に従うホワイトノイズがしばしば仮定されます。

正規分布を仮定する理由には様々ありますが、モデル化の容易さがまずはあげられます。

また、対数変換をするなどの処理によって、正規分布にデータを近づけることができることもあります。

2－6　時系列データの構造

時系列データの構造は、対象となるデータによって変わりますが、大きく以下の要素に分解して説明することができます。

時系列データ＝ 短期の自己相関
＋周期的変動
＋トレンド
＋外因性
＋ホワイトノイズ

もちろん、これらの要素が常にすべて入っているわけではなく、周期的変動がないデータや長期のトレンドがないデータなどもあります。

どの要素をどういう形式でモデルに組み込むかは、データ分析者の腕の見せ所といえるでしょう。

出版社 Web サイト：http://www.pleiades-publishing.co.jp/index.html
書籍のサポートページ：https://logics-of-blue.com/tsa-ssm-book-support/