時系列のEDA/シミュレーション

実践 時系列解析 輪読会（3章~4章）

Kazuki Igeta @kazukiigeta

2021/11/13

簡単に自己紹介

Kazuki Igeta; 井ケ田 一貴

Twitter: @kazukiigeta

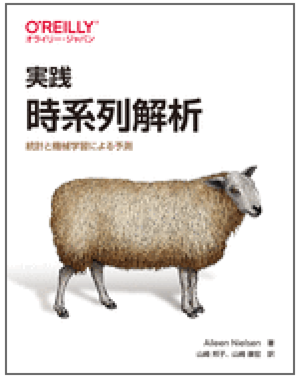
仕事

- ◆ 顧客企業向けの課題発掘＆AIで解決できる問題への変換
- ◆ 顧客企業に提供するAIモデル開発
- ◆ 社内DX（数理最適化、AI活用）など…

好きなもの

- つくること/学ぶこと全般
- サイバーセキュリティ
  - Mini Hardening運営メンバー（セキュリティ対策競技）
  - 顧客向けCTF作問 多数（セキュリティ関連競技）

本スライドについて



時系列データ解析について勉強するために
O’Reillyから出版されている**実践 時系列解析**を読み、
重要と感じた部分についてエッセンスを抜粋しました。

一部、他の書籍からの情報を追加しています。

細かい内容は**実践 時系列解析**の書籍を参照してください。

Agenda

- ➢ 時系列の探索的データ解析
- ➢ 時系列データのシミュレーション

3章 時系列の探索的データ解析

3.1 よく使用される手法

データ解析に際して
まず、**一般的な探索的データ解析（EDA）**の観点で以下を自問しよう
- 列の間に強い相関はあるか？
- 注目している変数の全体的な変数の全体的な平均値と分散はいくつか？

⇒【技法】プロット描画、要約統計量、ヒストグラム、散布図など

次に、**時系列データ解析**の観点で以下の自問をしよう
- 今見える値の範囲はどれくらい？期間や他の要素によって変化する？
- データが均質に測定されているように見える？時間的な変化が見える？

⇒【技法】EDAの技法に軸やグループとして時間を組み込む

3.2 時系列専用の探索的手法

- **定常性**の直観的な説明
  - 統計的性質（特に平均値と分散）が長期にわたって安定する性質
  - 従来の統計的時系列モデルの多くが定常性を仮定している

- 定常でない時系列
  - 定常であると断言するより定常でないことを示す方が簡単な場合もある
  - 例
    - 平均値が時間とともに増加
    - 分散が時間とともに増加
    - 強い季節変動を示す

3.2 時系列専用の探索的手法

- 定常過程
  - ある過程の取りうるすべてのラグkについて、 $y_t, y_{t+1}, \ldots, y_{t+k}$ の分布が t に依存しなければ、その過程は定常
  - 単位根の有無（過程の特性方程式の解の1つに1があるか否か）に帰着
    - 線形時系列は単位根があれば非定常（その逆は成り立たない）

- 単位根について
  - 単位根があると定常過程のように見えても非定常ということがあり得る
  - 例：$y_t = \varphi \times y_{t-1} + e_t$
    - $\varphi = 1$のとき、ランダムウォーク（単位根の代表例）
      - トレンドが内在しないが暴走して定常でないという良い例の1つ
    - $|\varphi| < 1$のとき、単位根なしの定常過程
      - 平均、分散が発散せず、各変数の分布がtに依存しない

3.2 時系列専用の探索的手法

- 拡張ディッキー･フラー（Augmented Dicky-Fuller; ADF）検定
  - 定常性の評価に最もよく使われる
  - DF検定の回帰モデル
    - $\Delta y_t = y_t - y_{t-1} = (\varphi - 1) \times y_{t-1} + \epsilon_t$
      - AR(1)モデル: $y_t = \varphi \times y_{t-1} + e_t$の両辺から$y_{t-1}$を引いたもの
      - y_{t-1}の係数が0かどうか、H_0: $\varphi = 1$の t 検定に置き換えられる
  - ADF
    - $y_t - \varphi_1 \times y_{t-1} - \varphi_2 \times y_{t-2} = \epsilon_t$
    - H_0: $\rho = \varphi_1 + \varphi_2 + \cdots + \varphi_2$=1の t 検定（ADF検定）
      - DF検定より多くのラグを考える

- 欠点
  - 単位根に近い根と単位根を区別する能力が弱い
  - 標本サイズが小さいとFPがよくある
  - 定常性そのものの検定ではない（単位根を持つか否かだけ）

3.2 時系列専用の探索的手法

- **定常でない**時系列モデル
  - 時系列の指標の変化に伴い精度が変化
  - モデルのバイアスと誤差が時間とともに変化
  - データの変換によって十分に定常となる場合が良くある
    - 分散が時間とともに変化するとき
      - 対数変換
      - 平方根変換
        - 上記変換の両方を試すべき
    - トレンドがあるとき
      - 階差（4回以上の階差は取るべきではない？）

3.2 時系列専用の探索的手法

- 時系列予測モデルにおける一般的な仮定
  - 定常性
  - 入力変数や予測変数の正規性
    - ボックス-コックス変換などで正規分布に近づけた方が良い場合もある
      - 変換が最も重要な情報を保存するように注意

3.2 時系列専用の探索的手法

- ローリングウィンドウ
  - 処理の対象とするデータの範囲を制限することで情報量を増やす方法の1つ
  - 左揃え
    - ウィンドウサイズ分の**過去**の値を対象とする
  - 右揃え
    - ウィンドウサイズ分の**未来**の値を対象とする
      - リークに注意
      - 事前に情報を持っていたら有用だったかの問いに役立つ
  - ドメイン知識があれば、カスタムローリング関数が適した場面もある

- 拡張ウィンドウ
  - ローリングウィンドウよりも利用シーンが限られる
    - 定常な時系列のみ
    - 情報を集めながらオンラインの要約統計量を維持するのに役立つ

3.2 時系列専用の探索的手法

- 自己相関
  - Rの`acf()`で確認可能
    - 自分で実装するよりも便利
  - 自己相関関数（ACF）の推定の棄却域は $\pm 1.96 \frac{1}{\sqrt{n}}$
    - 標本サイズが十分かつ分散が有限の場合のみ
  - 系列の和のACFは、個々のACFの和に等しい

3.2 時系列専用の探索的手法

- 偏自己相関
  - Rの`pacf()`で確認可能
  - ACFと違って、個々の系列の和のPACFは各系列のPACFの和より大きい
    - 周波数の異なる振動が続くため、近傍の点による影響が小さくなるため

3.2 時系列専用の探索的手法

- 共和分
  - 2つの時系列間の真の関係
  - 強い相関がみられる
- 見せかけの回帰
  - 共和分関係の存在しない2つの系列の間に強い相関がみられる場合が多々ある
    - ニコラスケイジの映画出演回数とプールでの溺死者
  - トレンドや季節性のある時系列同士でよくみられる
  - 強い相関を発見した場合には疑ってかかる必要がある
    - 例：ダービンワトソン統計量が0に近い場合は誤差項に自己相関がある 等

3.3 知っていると便利な可視化手法

- 1次元の可視化
  - ガントチャートを用いて、各系列の分布の重なりを捉えられる
    - Rの`timevis()`等

- 2次元の可視化
  - 年毎の可視化　：季節性を確認できる
    - Rの`seasonplot()`等
  - 月ごとの可視化：月ごとのトレンドの特性を確認できる
    - Rの`monthplot()`等

- 曜日、月、年などの時間軸でプロットすることで有用な情報が得られる

4章 時系列データのシミュレーション

書籍で紹介されるシミュレーション

- ヒューリスティックシミュレーション
  - 世界がどう動くかを決め、整合性を確保し、規則をコード化
    - 例：非営利団体の会員のメール開封と寄付行動
- 離散事象シミュレーション
  - 宇宙に特定の規則で動く個々のアクターを作り、動作させて宇宙の変化を観察
    - 例：シフト表が様々で集客頻度が時間依存するタクシー群の挙動
- 物理シミュレーション
  - 物理法則を適用して系の時間発展を観察
    - 例：ある磁性固体の特定条件における状態変化

- シミュレーションの活用
  - 本物のデータが得られる前にシミュレーションデータを使ってパイプラインを先んじて構築できる。

磁性個体の状態変化シミュレーション例で使われた理論

- MCMC（Markov chain Monte Carlo）
  - モンテカルロシミュレーション
    - 理論上は厳密買いがあるが確率的に解く方が簡単な場合にそうする方法
    - 項の特定の分布や級数の様子が分かる
  - マルコフ連鎖
    - 次の状態が現在の状態に遷移する確率は現在の状態にのみ依存
    - モンテカルロシミュレーションで分かった様子の時間変化が分かる

- MCMCの１つであるメトロポリス法

まとめ

- Rをつかった探索的データ解析を実行する手順
  - 時系列データを見たら必ず実行するような初歩的な部分について学ぶことができる
- Pythonをつかった時系列データシミュレーション
  - 現実世界の現象をルールとして羅列し、仮定する確率分布などの適切なロジックを考え、コードに落とし込み、データを生成するという流れを体験できる