初心者のための

Python 時系列解析 入門

https://to-kei.net

全人類がわかる統計学 Presents

本日の目標

Pythonで時系列解析の流れに触れること

目次

- ➢回帰分析について
- ➢時系列解析について
- ➢時系列データの前処理
  - 定常、非定常、自己相関係数、偏自己相関係数
- ➢時系列モデル
  - 時系列モデル説明と実装
    AR、MA、ARMA、ARIMA、SARIMAモデルの実装
- ➢時系列モデル選択と予測精度
  - AIC、RMSEの説明と実装
- ➢まとめ

回帰分析について

回帰分析とは

- ある値(X)とある値(Y)がなにか関係がありそうだ。
  これらの値を使ってYの値が予測できるのではないか？？

  という発想からXとYを

$$Y = \ f(X)$$

という数式に当てはめて予測をする手法である

体重(X)と身長(Y)というデータから身長を予測する場合

体重 (X)	身長 (Y)
65	175
70	185
40	165
80	180
55	170
60	180

$$\mathrm{Y} = f(X) = aX + b$$

というモデル式で予測を行う。

このとき、

X：説明変数。Yの予測に使う変数
Y：目的変数。予測される側の変数
A：回帰係数
B：切片

と呼ぶ

最小二乗法

未知のパラメータaとbを求めるために、最小二乗法という手法を用いる

最小二乗法の考え方

正確なaとbを求めるのは難しい…
そこで、

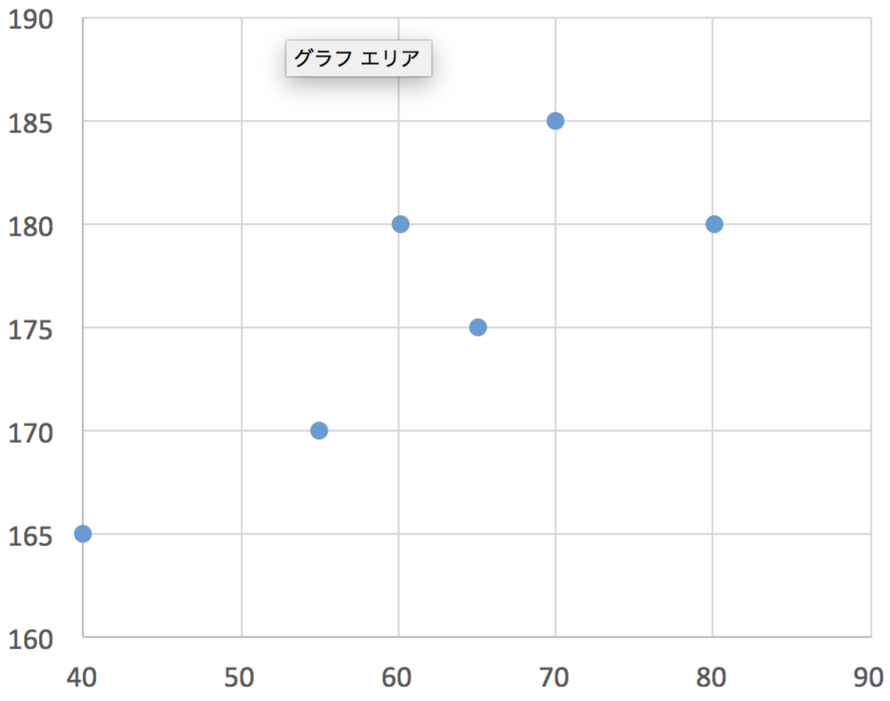
$$Y = f(X) = aX + \mathrm{b} + \varepsilon$$

新たに誤差(ε) があると考え、
この誤差を最小にするような値を求める

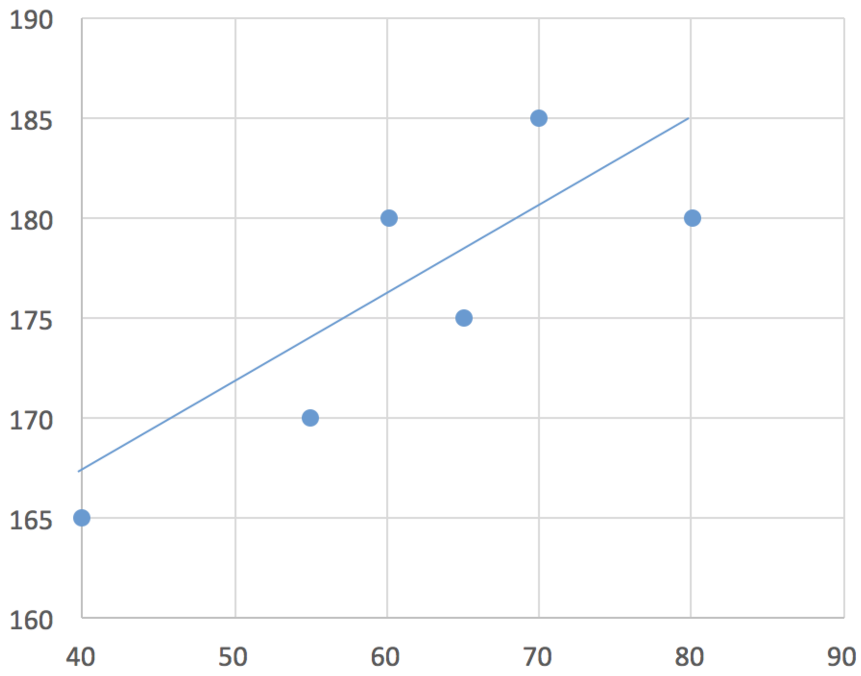
最小二乗法の直感的理解①

体重 (X)	身長 (Y)
65	175
70	185
40	165
80	180
55	170
60	180

データ

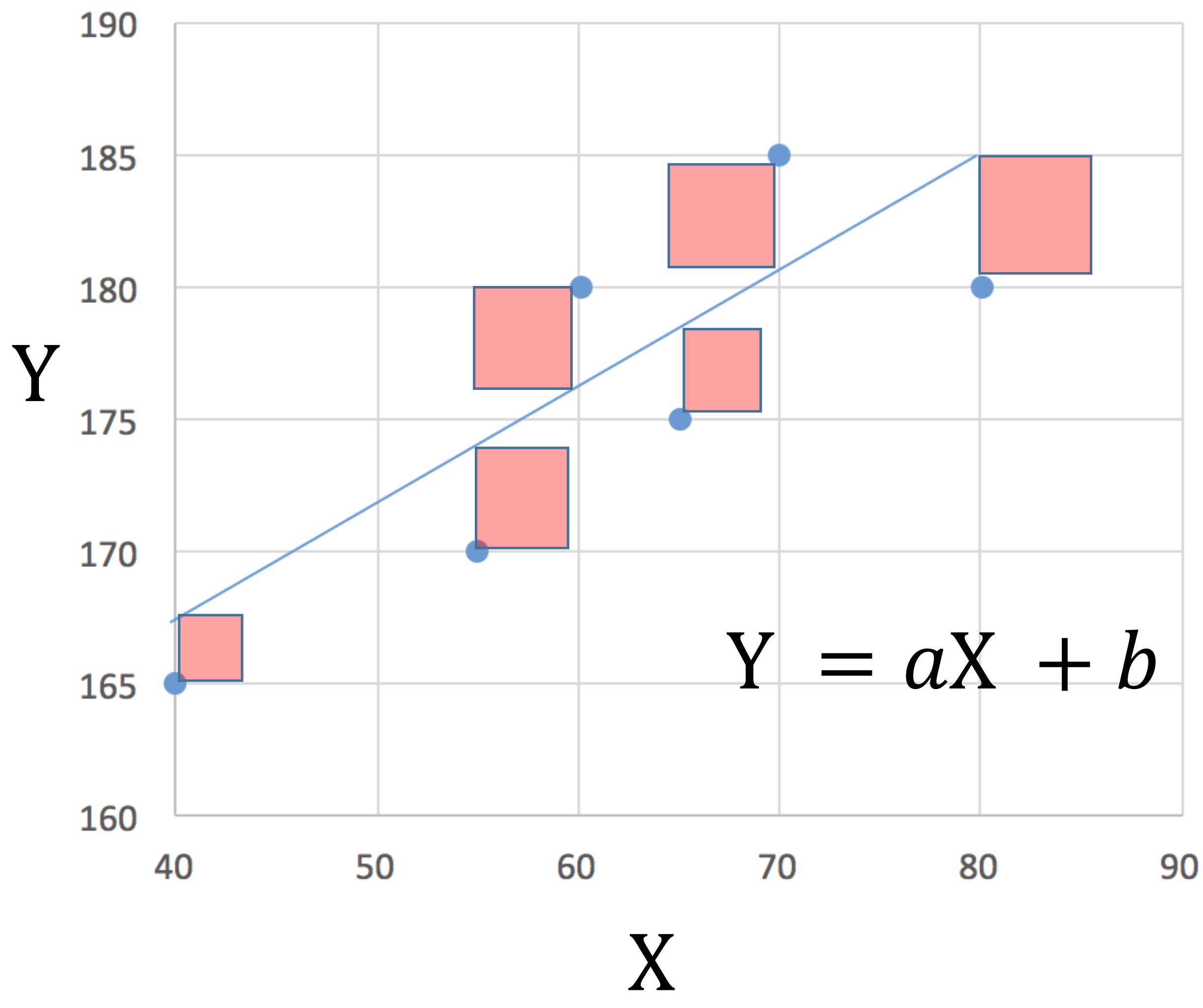


散布図



回帰直線

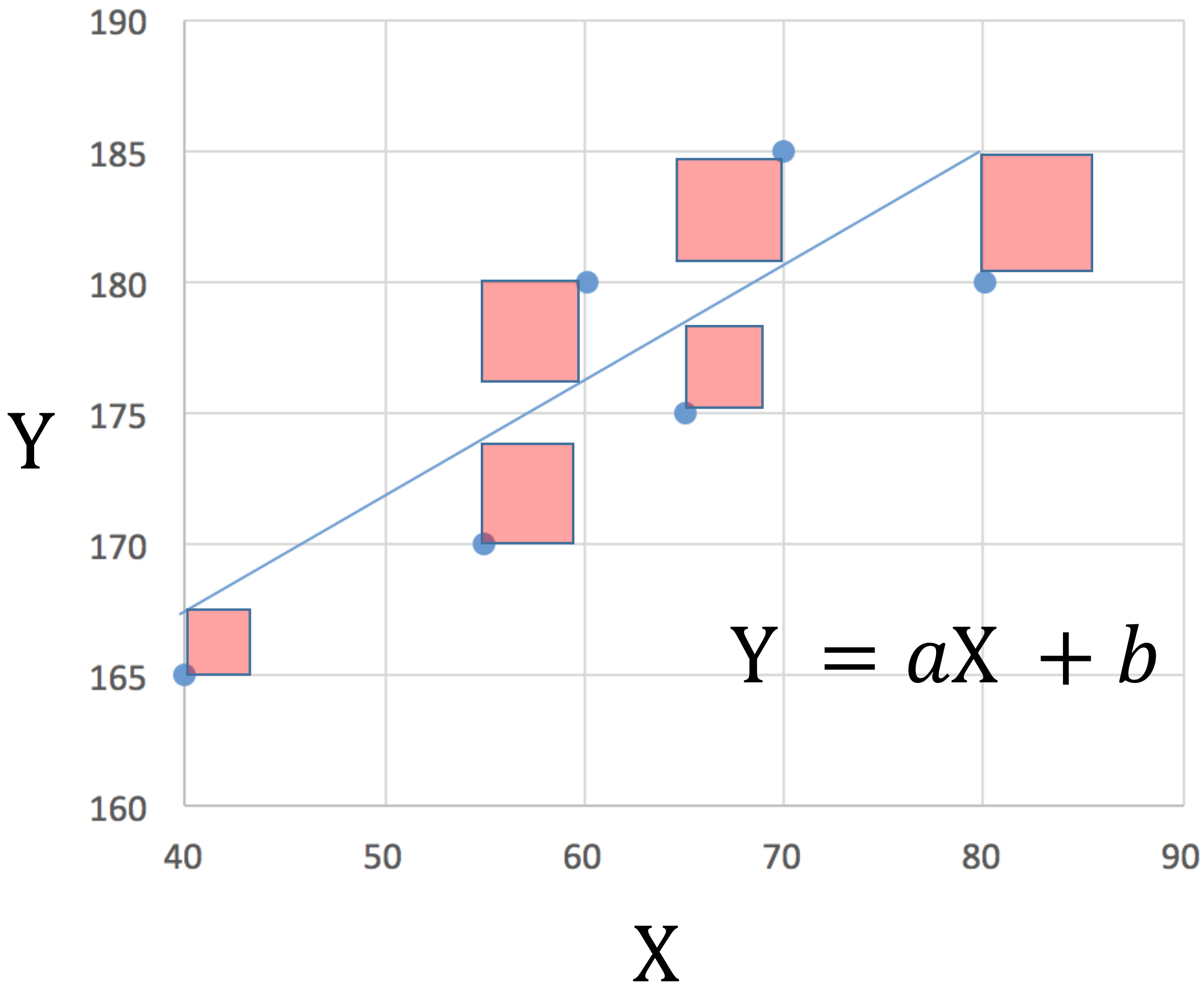
最小二乗法の直感的理解②



$$\varepsilon^2 = \sum_{i=1}^{n} (Y_i - aX_i + b)^2$$

εの値のままだと誤差が正負で打ち消しあってしまうので二乗した面積(赤色の正方形)を考える

最小二乗法の解



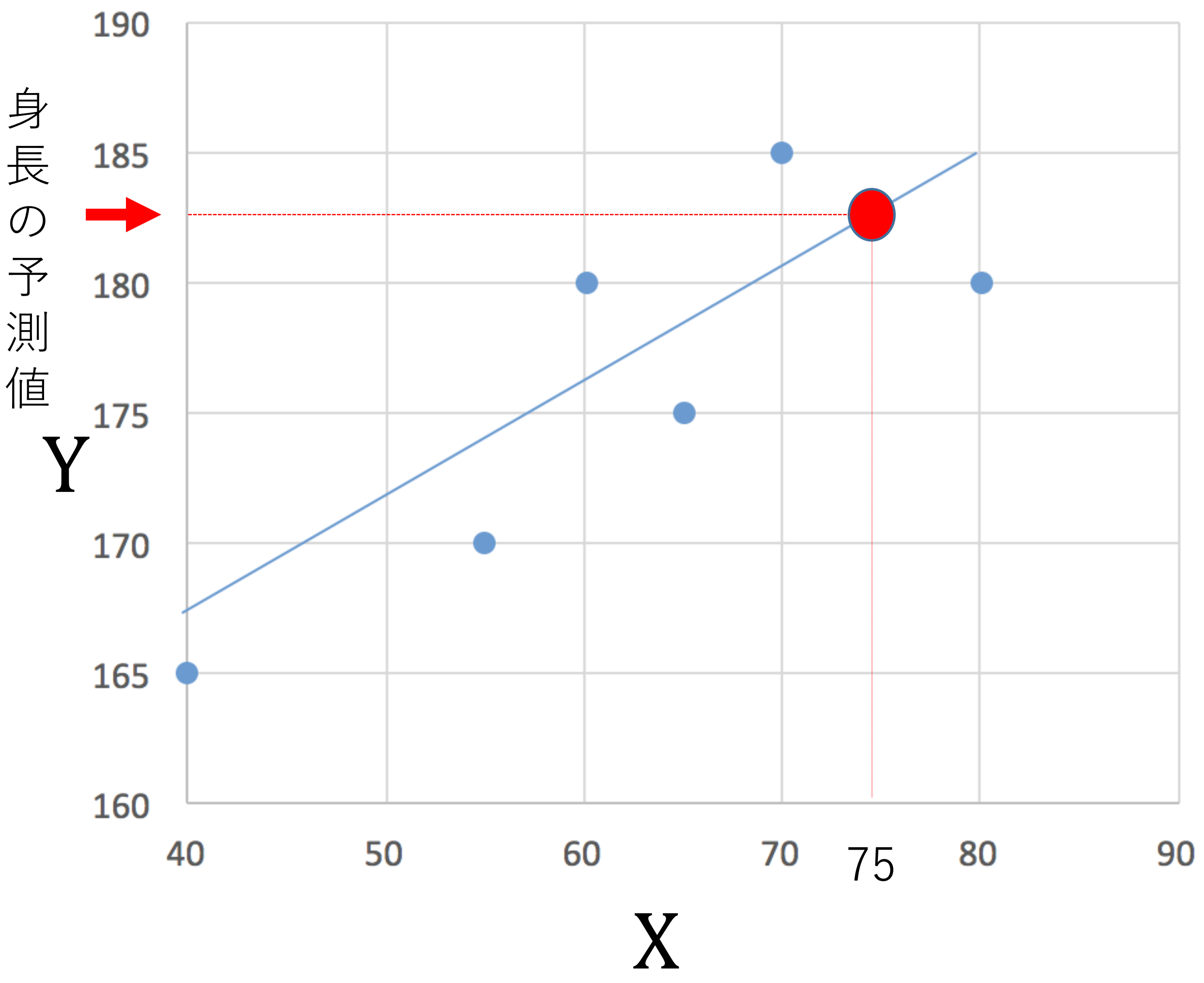
赤色の面積を最小すること

$$\varepsilon^2 = \sum_{i=1}^{n} (Y_i - aX_i + b)^2$$

この式からa, bを求めること
(偏微分して=0)

回帰係数 a と切片 b は過去のデータを使用して求められる！

予測するときは？



新しいデータの体重**75kg**の人の身長は

$$Y = a \times 75 + b$$

で予測できる。

過去のデータから予測

→ **教師あり学習**

回帰分析のまとめ

- Yに、$Y = f(X) = AX + B$というモデルを仮定して予測
- Xが1つの場合は単回帰分析、複数の場合は重回帰分析
- **回帰分析は機械学習の教師あり学習**

時系列解析について

時系列解析とは？

時間の流れとともに観測されたデータを解析して、変化の規則などを見出す解析手法

- Googleトレンド、為替レート、GNP、ある地点での気温、etc

時系列解析の回帰分析との相違点

時系列データは

「毎日の売り上げデータ」「日々の気温のデータ」
*「月ごとの飛行機乗客数」　「四半期決算データ」*など

毎日（あるいは毎週・毎月・毎年）増えていくデータ

- ➢データ間にはある程度の相関があると考える
- ➢一般的な回帰分析を行うと見せかけの回帰に陥る

⇨ **時系列データに対しては時系列解析を行おう！**

代表的な時系列解析の手法分類

時系列解析

- *機械学習的アプローチ*
  - *統計的アプローチ*
    - **状態空間モデル**
      - **(非)定常時系列モデル**
  - Randomforest
  - Prophet
  - など
  - *深層学習的アプローチ*
    - RNN
    - LSTM
    - など

時系列データの前処理

~データの読み込み~

時系列データの読み込み例①

```
import pandas as pd
pd.read_csv("XXXXX.csv")
```

	Month	#Passengers
0	1949/1/1	112
1	1949/2/1	118
2	1949/3/1	132
3	1949/4/1	129
4	1949/5/1	121

➢pandasのメソッド .read_csv()を用いる

➢ファイルのデータがDataFrame型で読み込まれる

※読み込ませるcsvファイルは作業ファイルと同一フォルダ内に配置(絶対パスで入力も可)

時系列データの読み込み例②

```
pd.read_csv('XXXXX.csv', index_col='XXXX ', parse_dates=True, dtype=float)
```

	#Passengers
Month	
1949-01-01	112.0
1949-02-01	118.0
1949-03-01	132.0
1949-04-01	129.0
1949-05-01	121.0

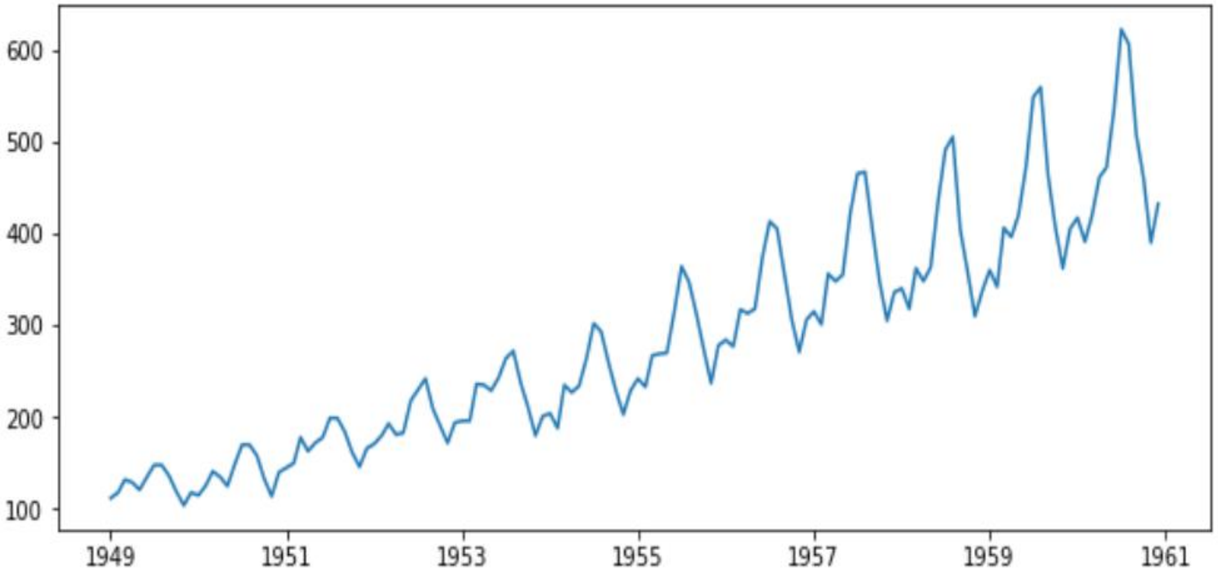
- index列を'XXXX'の列とし、DataFrame型変数からデータを切り取っても日付のindexを取得できるようにする
- 時系列解析をする場合はデータをfloat型として読み込む

時系列データの前処理

~データの可視化~

時系列データの描画

```python
from matplotlib import pyplot as plt
plt.figure(figsize=(12, 4))
plt.plot(XXXXX)
plt.show()
```



- .figure(figsize=(,))で描画するキャンバスの大きさを設定
- プロットしたいデータを引数として .plot() で描画
- .show() で確認できる

※ jupyter notebookの場合は記載しなくとも表示される

時系列データの前処理

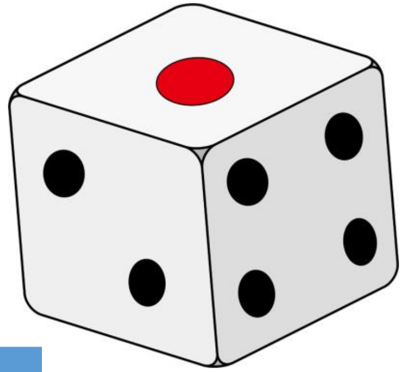
~時系列データの性質~

時系列データの性質

- ➢時系列データは
確率過程と呼ばれる時間に依存した確率変数列からの実現値である

時系列データの性質の直感的理解①

例えば、1~6の目が同じ確率で出現するサイコロがある。



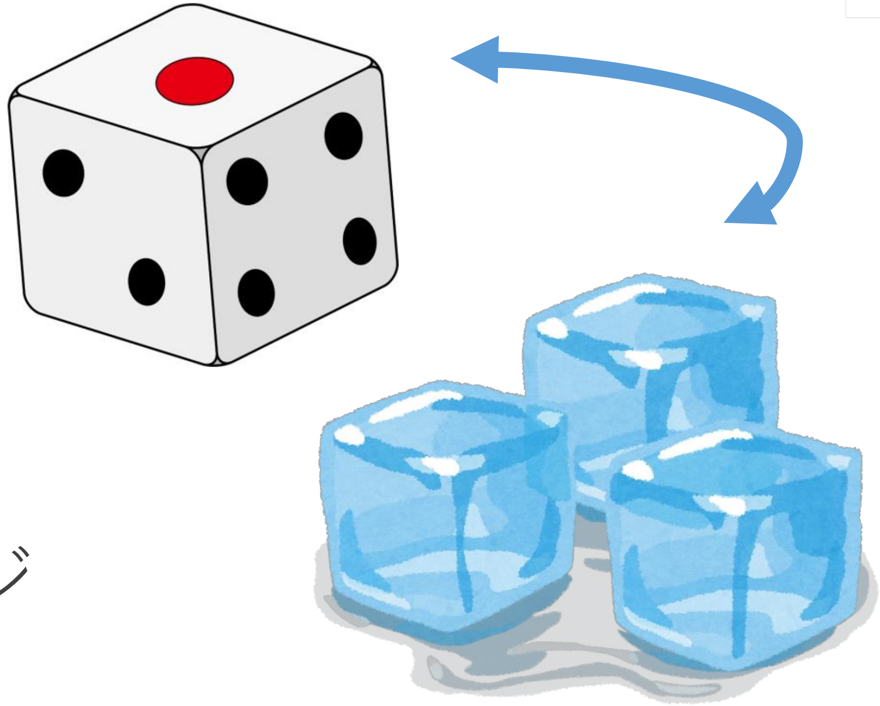
確率変数	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

※確率変数とはサイコロを振って出る目のこと

時系列データの性質の直感的理解②

時系列の場合、

サイコロの材質が氷できている



というイメージ

時間が経過すると氷が解けて1~6が出る確率が変化

時系列データの性質の論理的理解(サイコロの例)

3

4

時系列データ列

$$y = \{y_1, y_2, \cdots \cdots, y_t\}$$

時間の経過

$$Y = \{Y_1, Y_2, \cdots \cdots, Y_t\}$$

確率変数列

1,2,3,4,5,6

2,4,6

時系列データの性質の論理的理解(サイコロの例)

Y_1	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

Y_2	1	2	3	4	5	6
確率	0	1/3	0	1/3	0	1/3

時系列データの性質の論理的理解

時系列データ列

$$y = \{y_1, y_2, \cdots \cdots, y_t\}$$

時間の経過

$$Y = \{Y_1, Y_2, \cdots \cdots, Y_t\}$$

確率変数列

時間の経過で確率変数が変わってしまう…

時系列データの性質の論理的理解(サイコロの例)

確率変数が時間で変化する上にサンプルは1つしか得られない…

3

4

$$y = \{y_1, y_2, \cdots \cdots, y_t\}$$

時間の経過

$$Y = \{Y_1, Y_2, \cdots \cdots, Y_t\}$$

1,2,3,4,5,6

2,4,6

そこで

定常時系列モデルという枠組みでは
データに***(弱)定常性***という仮定をおいて
解析を可能にした

定常時系列モデル

時系列解析

機械学習的アプローチ

統計的アプローチ

状態空間モデル

定常時系列モデル

Randomforest
Prophet
など

深層学習的アプローチ

RNN
LSTM
など

（弱）定常性について

定義

任意の時点 t とラグ k に対して

$$E(y_t) = \mu$$
$$Cov(y_t, y_{t-k}) = E[(y_t - \mu)(y_{t-k} - \mu)] = \gamma_k$$

- **観測されたデータが時間によらず、期待値と分散が一定**
- **自己共分散が時点ではなくラグに対して依存**

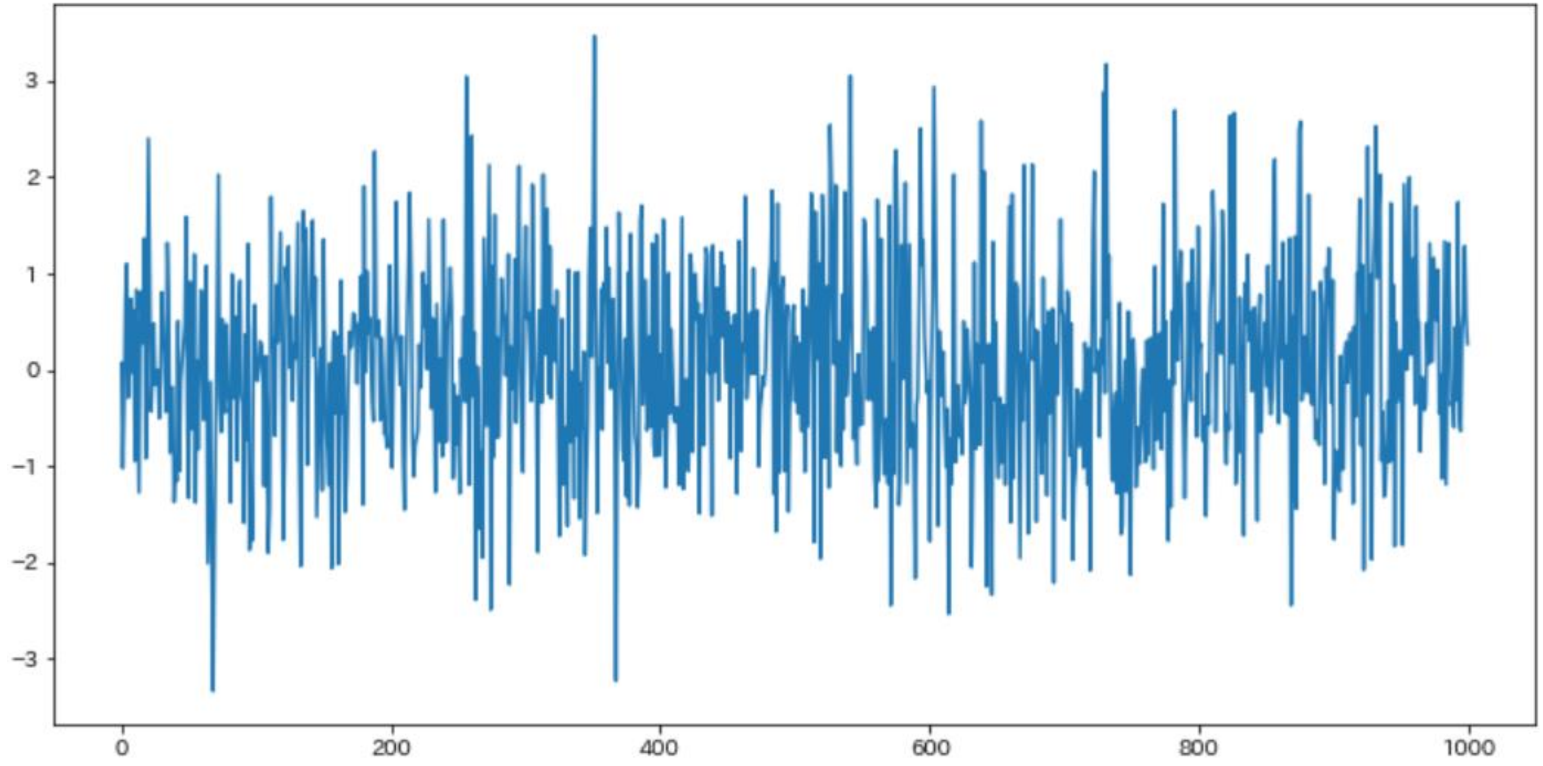
※自己共分散とは、元のデータとそれを何地点かずらしたデータとの関係性を示す指標です。

（弱）定常性について

(例)

$$E(y_t) = 0$$
$$Var(y_t) = 1$$

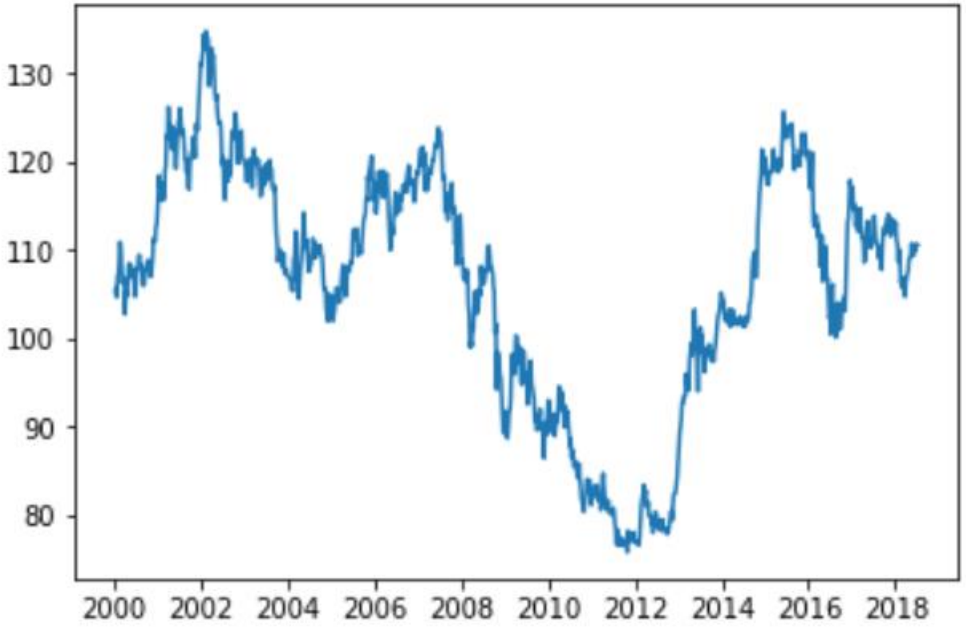
のデータ1000個

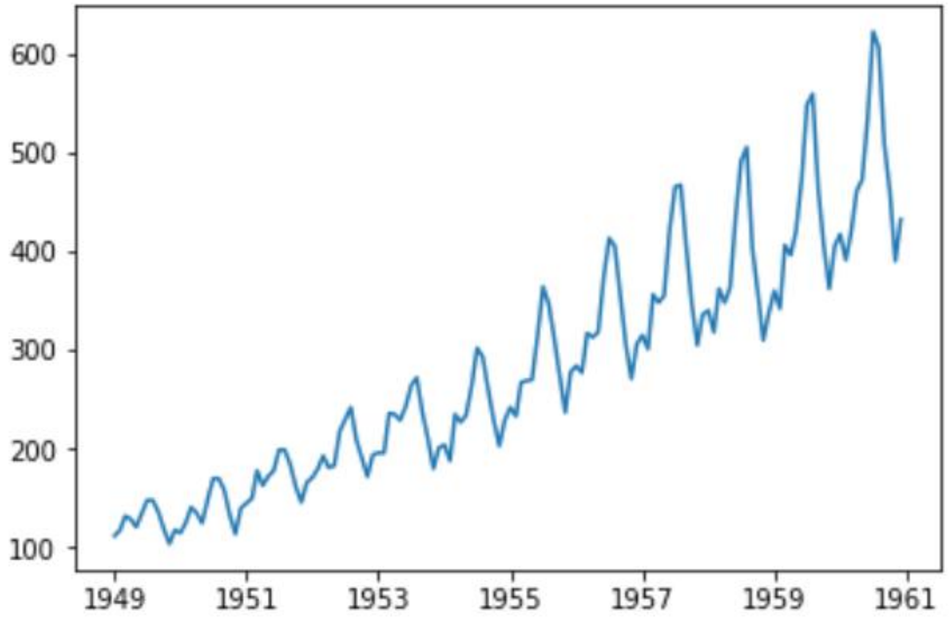


観測されたデータの期待値と自己共分散が一致するので
一定の値の幅で振動するような図となる！！

非定常性な時系列データ

現実に扱う時系列データは非定常な時系列データがほとんどである





非定常な時系列データは弱定常性に変換して解析する！！

非定常性な時系列データの変換

- **対数変換**
- **階差変換**
- **季節調整変換**

など様々な変換方法がある

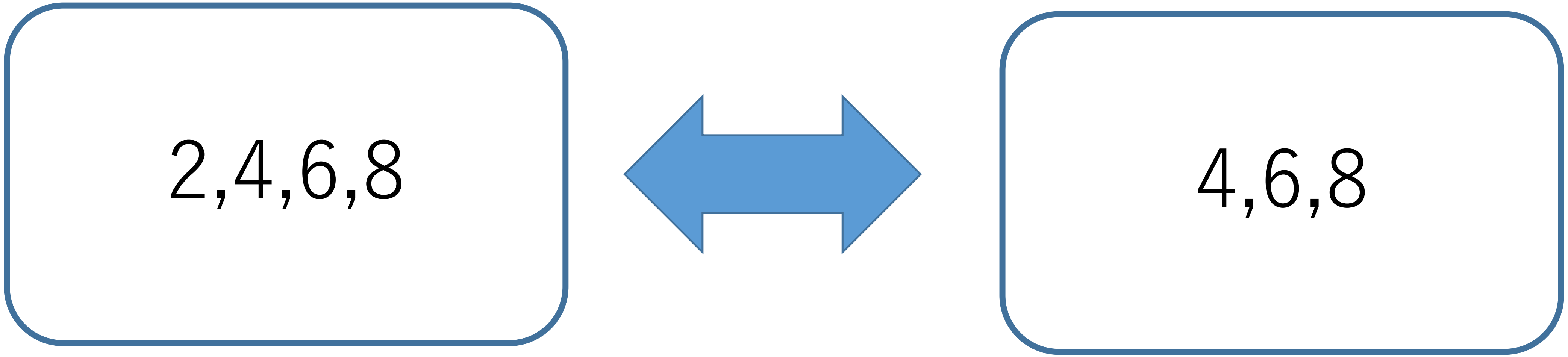
時系列データの種類まとめ

原系列	何も加工していない時系列データ
対数系列	原系列に対数変化したもの
階差系列/差分系列	原系列の各点から何時点か前の時点を引いた系列
対数差分系列	原系列に対数変換を施して、その系列の差分系列をとったもの
季節調整済み系列	季節変動の影響を取り除いた系列

時系列データの前処理

~データ間の関係性~

自己相関係数(-1~1)



- データ間の関係性を表す指標
- **1だと100%正の相関、-1だと100%負の相関がある**
- 自己共分散を標準化したもの

自己相関係数(-1~1)

r_k:自己相関係数

観測データの任意の時点 t とラグ k に対して

$$\mathrm{r}_k = \frac{\Sigma_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\Sigma_{t=1}^{n}(y_t - \bar{y})^2}$$

例えば、
1日前と大きな正の自己相関があれば、
「1日前に多ければ、今日も多い」ということになり、

2日前と負の自己相関があれば、
「2日前に多ければ、 今日は少ない」などと判断する

自己相関係数の数値計算

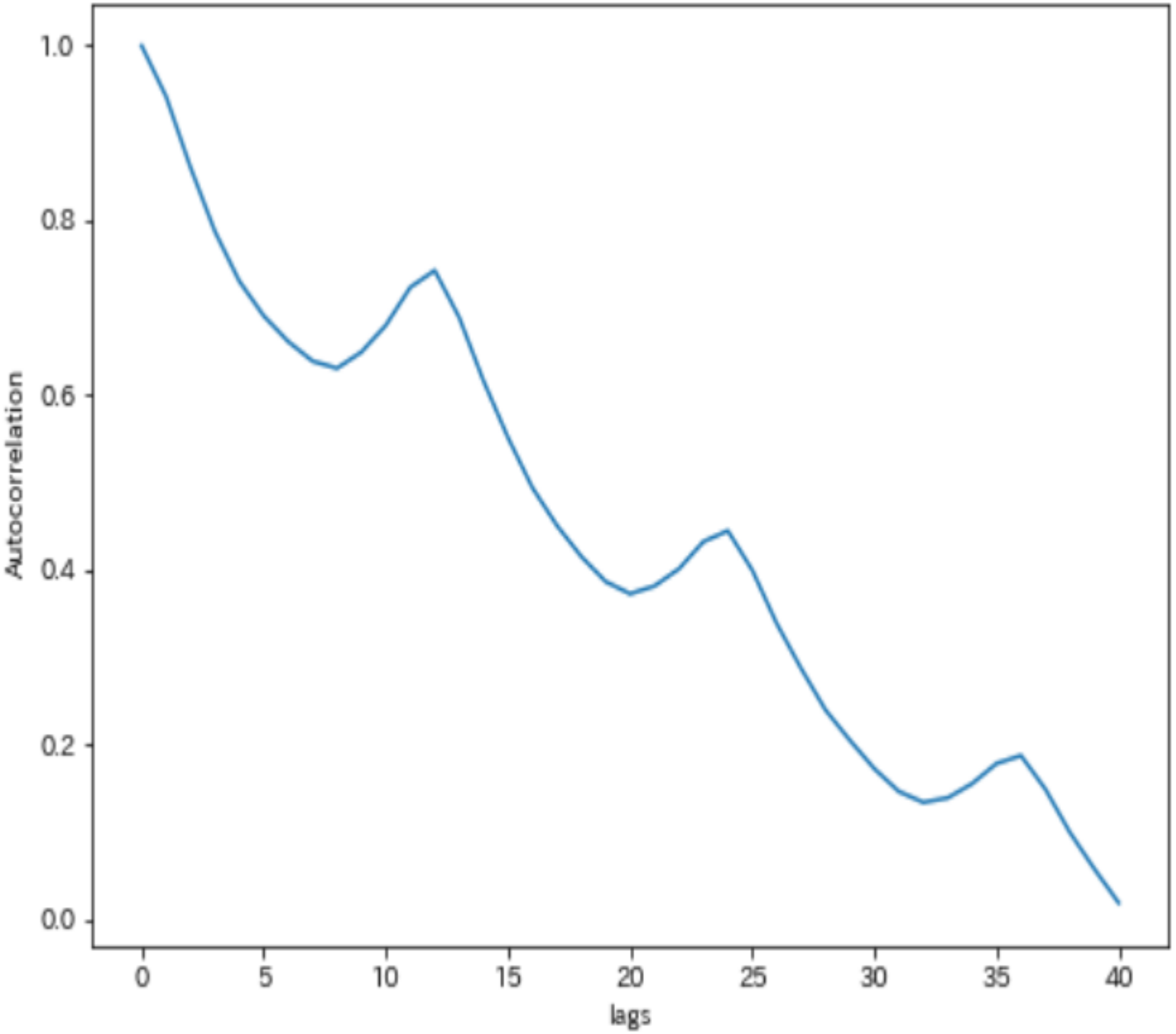
1次の自己相関

$$y_0 = 4\ 3\ 7\ 4\ 9$$
$$y_1 = 4\ 3\ 7\ 4\ 9$$

2次の自己相関

$$y_0 = 4\ 3\ 7\ 4\ 9$$
$$y_2 = 4\ 3\ 7\ 4\ 9$$

自己相関係数の読み取り



- 左図の場合、正の相関をもつ
- ラグが大きくなると係数が減衰していき、関係が消滅する
- 自己相関係数がある整数倍で大きくなっているような場合には周期性を考える

偏自己相関係数(-1~1)

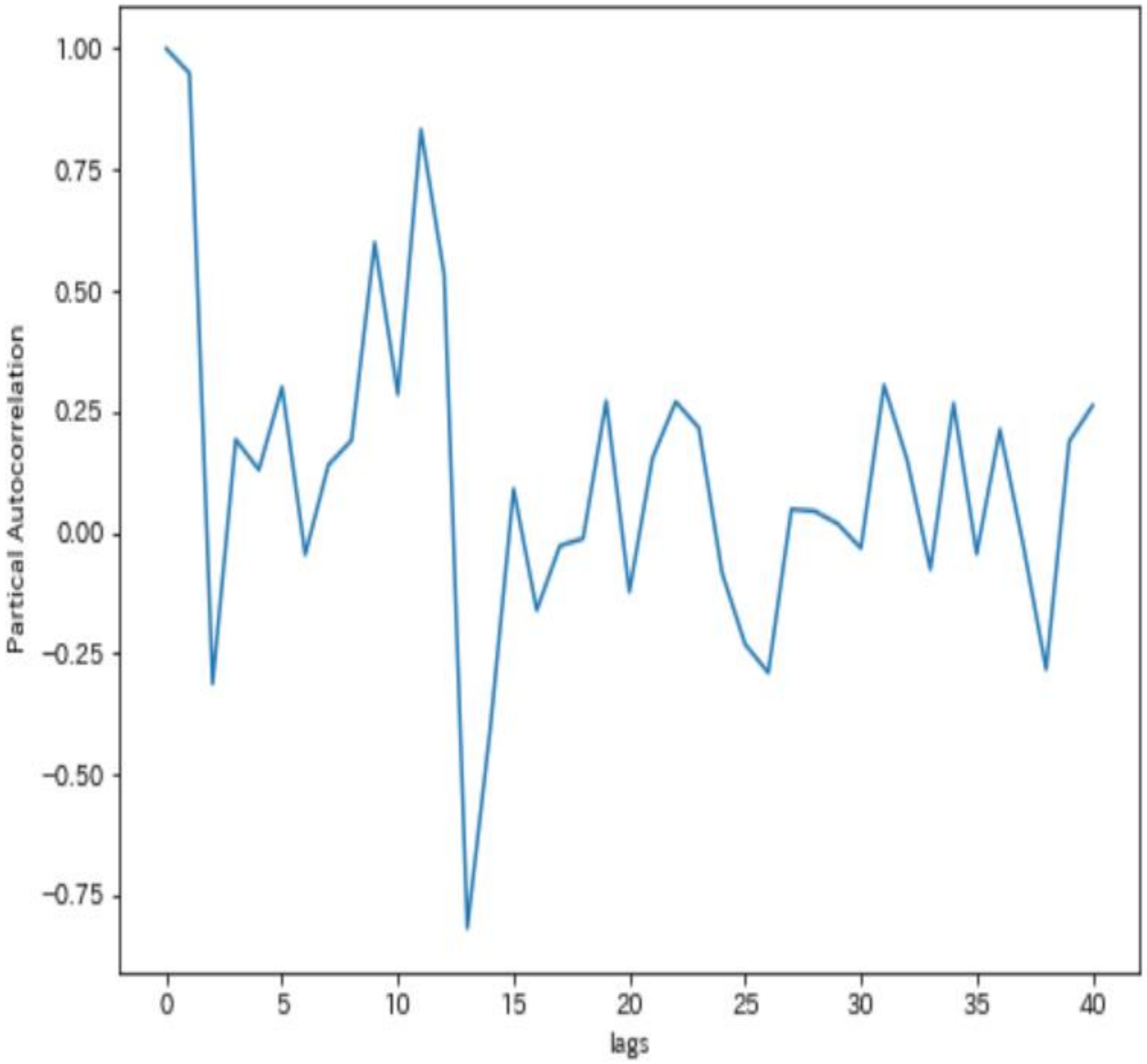
観測データの任意の時点tとラグk に対して

注目しているデータの観測点以外の要因を無視して計算された
自己相関係数(単純には求められないので式は省略)

偏相関係数を見れば、
ラグk(時点からのずれ)との
絶対的な関係を知ることができる

例えば、
一日前と今日の関係だけが知りたいときに用いる。
自己相関係数では一日前のデータが一昨日のデータの影響などを受けていると考える。

偏自己相関係数の読み取り

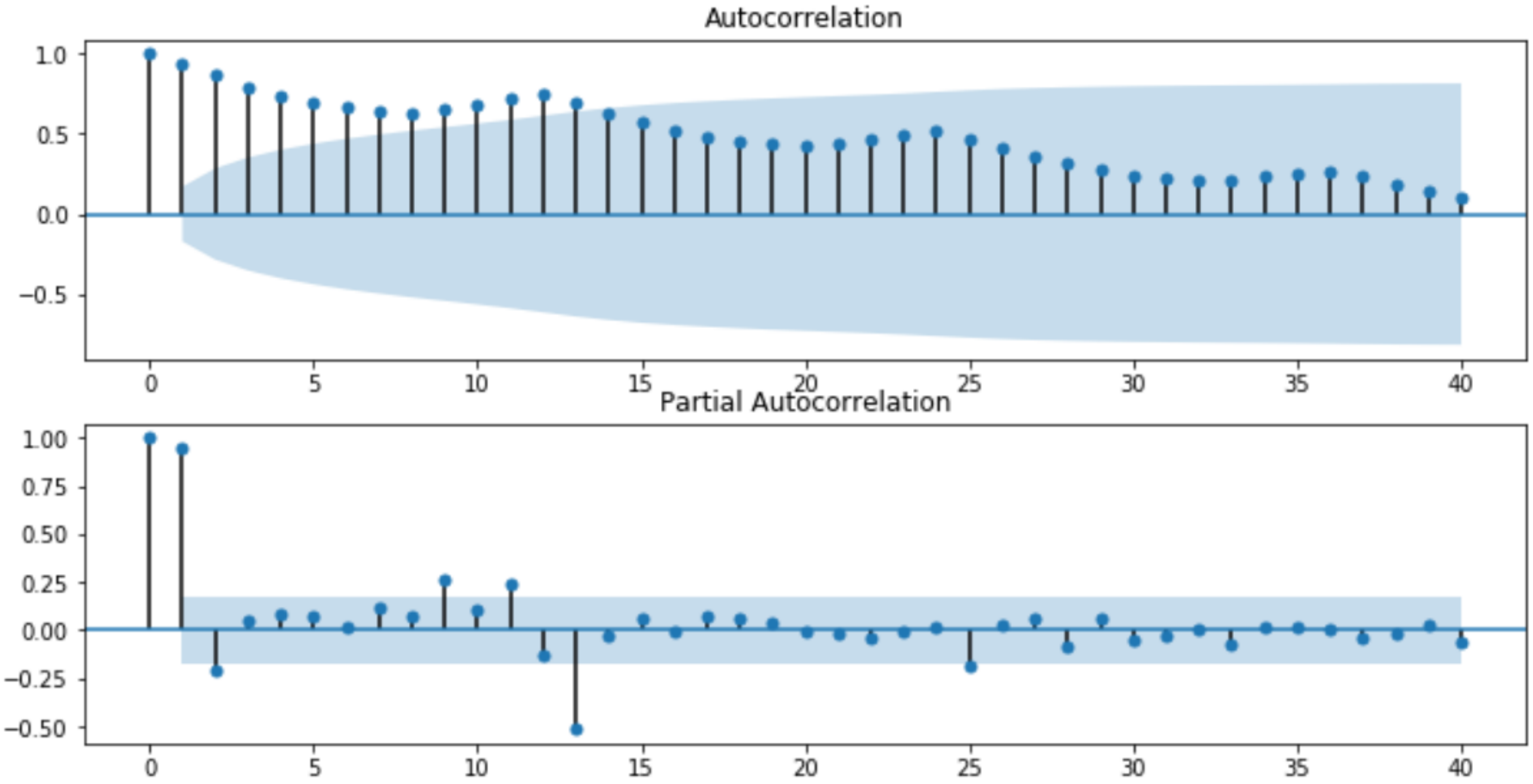


自己相関係数では見られなかった負の相関をもっているラグを発見することができた

※左図は自己相関係数のデータと同様のものを使用

信頼区間付き（偏）自己相関係数の見方

水色の領域内から飛び出すと相関があるとみなしてよい



問題

配布したデータ”AirPassengers.csv”を用いて以下の問題を実装せよ

① csvファイルをDataFrame型の変数として読み込み、データの数を確認せよ

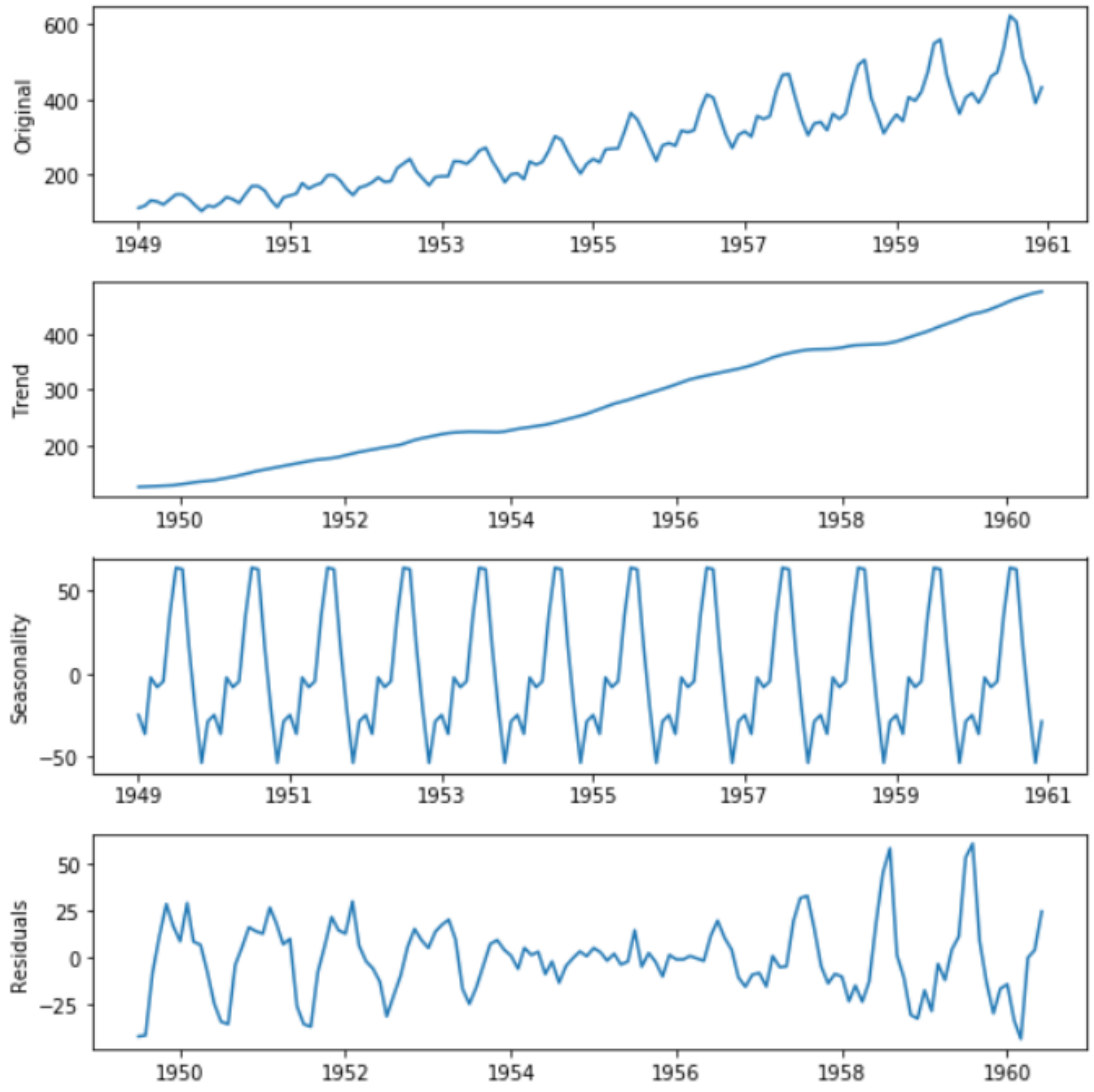
② データを可視化し、定常性があるか確認せよ

③ データの偏自己相関係数を求め、データを可視化せよ

時系列解析について

〜時系列モデル〜

時系列モデルとは?



時系列データをある仮定に基づいたモデルとして扱うこと！

モデルを作成できれば未来の値を予測するのに非常に便利！！

今回は古典的な
5種類のモデルを取り扱う

ARモデルについて

$AR(p)$の定義

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t$$

y_tと$y_{t-1}, y_{t-2}, , , , y_{t-p}$を用いて$y_t$を表現
p期前までのデータを使う場合には、$AR(p)$と表現

- *ε は誤差項で、平均0,分散σ^2 のホワイトノイズ（弱定常）*
- 基本的にARモデルにおけるϕは$|\phi| < 1$で与える
  （$|\phi| > 1$では発散して非定常となる）
- **p の値はデータの偏自己相関係数の図より判断**

ホワイトノイズについて

- データの発生には統計的にばらつき、つまりは誤差が出現
- 時系列データではこの誤差のことをホワイトノイズとして扱う

$\boxed{\text{平均が0, 分散が}\sigma^2 \quad \text{自己共分散が0}}$

平均が0, 分散がσ^2

自己共分散が0

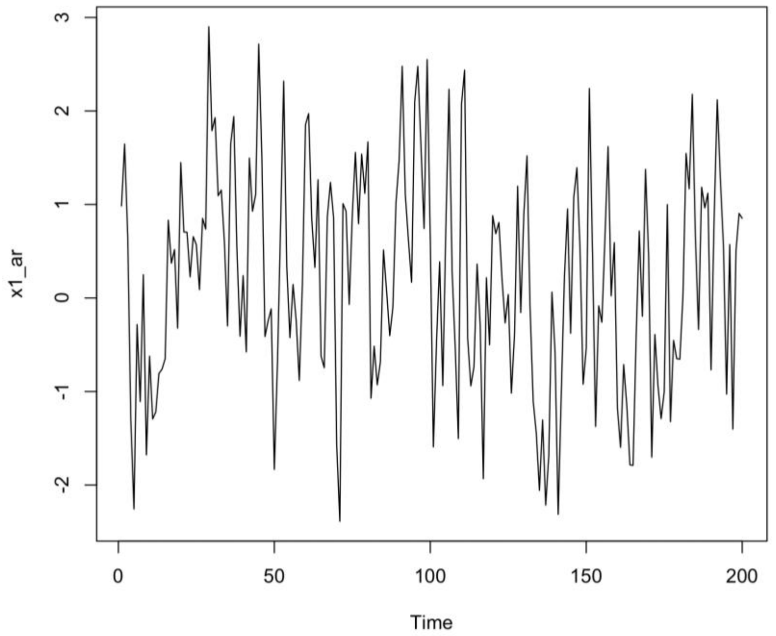
→

弱定常性を満たす

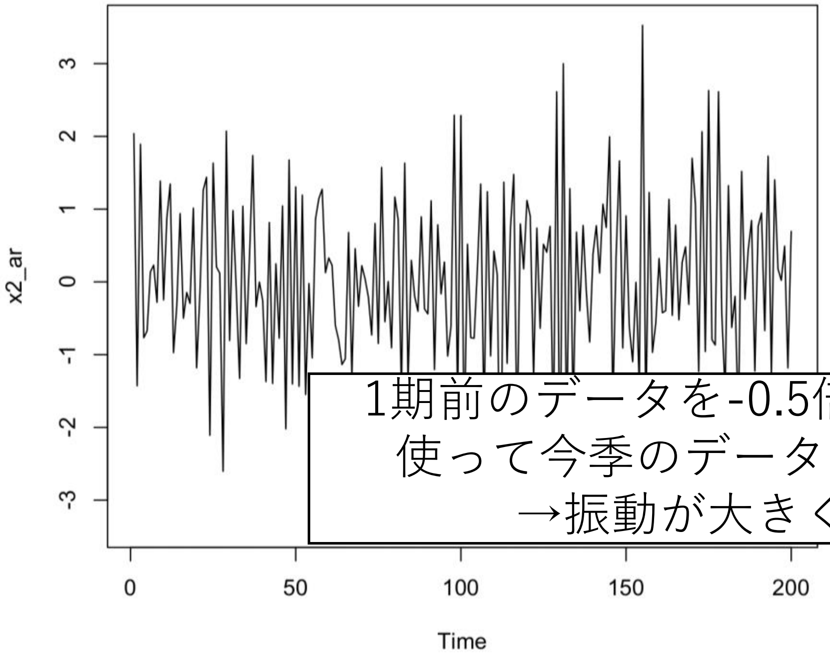
$AR(1)$モデル

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$

1つ前のデータと経験的に学習したパラメータ(cとϕ_1)と
ホワイトノイズ(誤差)を考慮して予測



$c = 0, \phi_1 = 0.5$の$AR(1)$のシミュレーション



1期前のデータを-0.5倍したものを
使って今季のデータをプロット
→振動が大きくなる

$c = 0, \phi_1 = -0.5$の$AR(1)$のシミュレーション

$AR(1)$モデルと単回帰モデルの比較

AR(1)モデル

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$

ϕ_1 ： 係数

c ： 切片

ε_t ： 誤差

単回帰モデル

$$y = b + a_1 x + \varepsilon_t$$

a_1 : 回帰係数

b ： 切片

ε ： 誤差

MAモデル

$$MA(q)$$

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

q期前までのデータの誤差の和を用いて表すモデル
この場合のモデルは$MA(q)$と表現する

- **過去のノイズが大きかった場合は、現在の値も**(θの重み付けを受けるものの)**大きく変化**する

ARMAモデル

$$ARMA(p,q)$$

$$y_t = c + \sum_{i=i}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

p期前までのデータとq期前までの誤差の和を用いて現在のデータを表現するモデル

- $ARMA(p,0) = AR(p)$
- $ARMA(0,q) = MA(q)$
- **ARMAモデルはARモデルとMAモデルの組み合わせ**

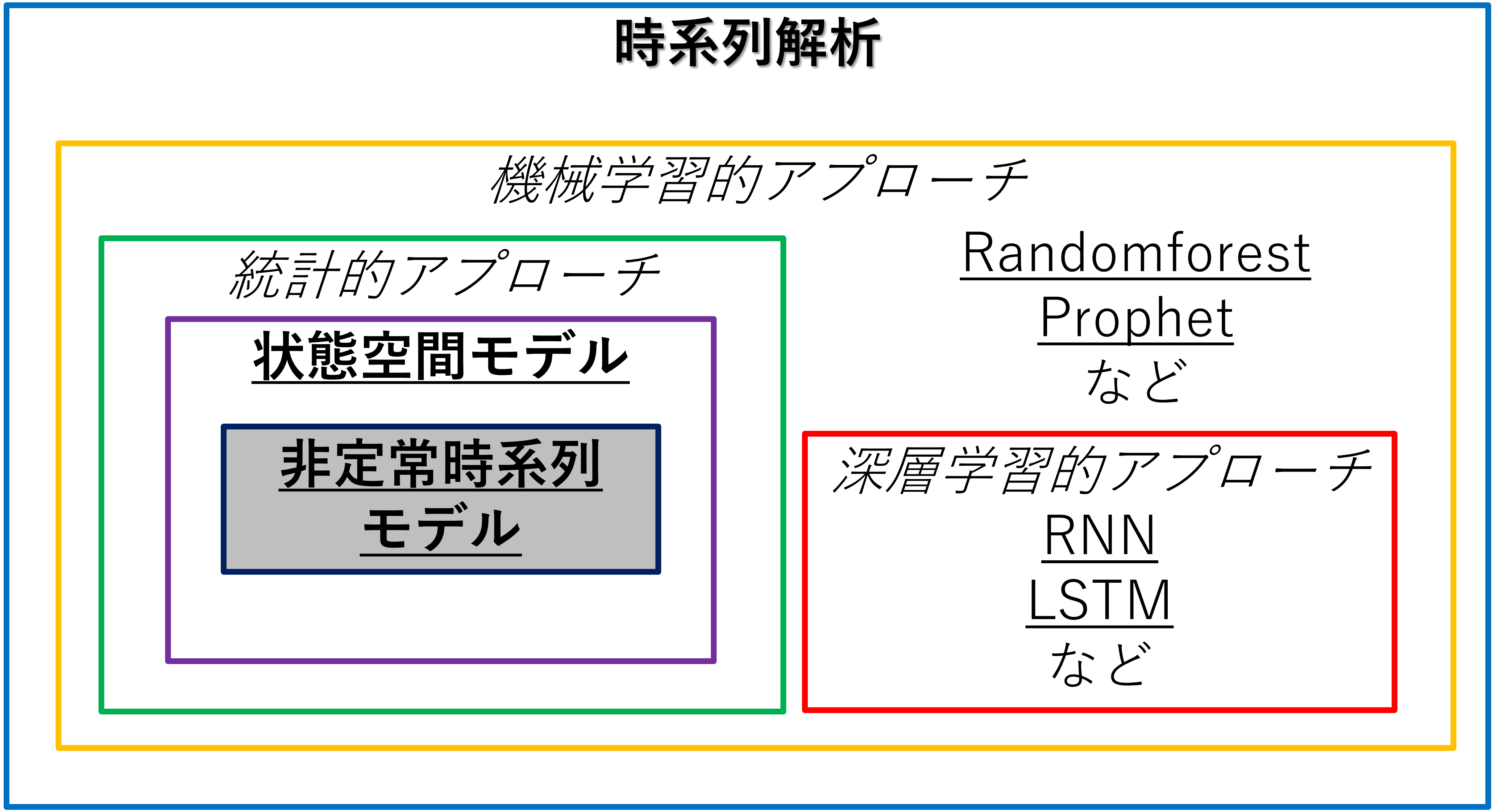
非定常時系列モデル

時系列解析の中の定常時系列モデルの枠組みでは
データに***(弱)定常性***が仮定できないと解析が行えない

↓

非定常な時系列データも扱いたい！！！

非定常時系列モデル


時系列解析
機械学習的アプローチ
統計的アプローチ
状態空間モデル
非定常時系列
モデル
Randomforest
Prophet
など
深層学習的アプローチ
RNN
LSTM
など

ARIMAモデル($ARIMA(p,d,q)$)

- ARMAモデルをさらに拡張し、学習データに対して何度も差分をとることで非定常なデータに対しても予測を可能にしたモデル
- 階差(d)は1~2ぐらいが基本

季節性ARIMAモデルについて（$SARIMA(p,d,q,sp,sd,sq)$）

- ARIMAモデルに、さらに周期的な変動（季節変動とか）を考慮した拡張モデル
- 季節性のあるデータに対して有効なモデル
- 季節性のパラメータ(sp,sd,sq)は低めに設定する

問題

配布したデータ”AirPassengers.csv”を用いて以下の問題を実装せよ

④ データに周期性があるか確認せよ。また、確認できる場合はどれくらいの周期か考察せよ

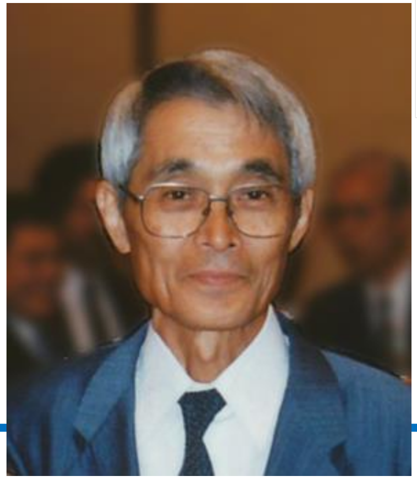
⑤ 今日学んだ時系列モデルの枠組みだと、どのモデルが一番有効であるか考察せよ

時系列モデルの選択と予測精度

～AIC～

AICについて

AIC（赤池情報量基準）について



➢良いモデルを選択するのに用いられる値の指標

値が最小となるものを予測精度が高いモデル
値は相対値となるので注意（関係のないモデルとは比較できない）

$AIC = -2 *$ (最大化対数尤度) $+ 2 *$ (推定されたパラメータ数)
で与えられる

情報量基準には他にも、BICなど様々なものがある…

RMSEについて

RMSEについて

- 予測精度を調べるための指標
- 求める式は以下で与えられる

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

- **値が小さいほど、実データと推定値が一致していると判断**

他にも、MAE、MSEなど様々な指標がある…

問題

配布したデータ”AirPassengers.csv”を用いて以下の問題を実装せよ

⑥ 講義のプログラムを参考にして、SARIMAモデルの中で当てはまりのよさそうな次数をAICを用いて求めよ

まとめ

今後学ぶべき解析手法

時系列解析

機械学習的アプローチ

統計的アプローチ

状態空間モデル

(非)定常時系列モデル

Randomforest
Prophet
など

深層学習的アプローチ

RNN
LSTM
など

主催講座一覧(1)

- 統計学入門講座
  - 超入門（ゼロ～統計検定3級合格レベル）
  - 確率変数・確率分布編
  - 推定・仮説検定編
  - ベイズ・回帰・分割表解析編

  $\}$ 統計検定2級

- Python機械学習・データ分析編
  - Python超入門
  - Python入門
  - Python文法演習
  - Pythonデータ分析入門
  - Pythonスクレイピング入門
  - Python機械学習入門
  - 実用的な機械学習モデルを作る演習

- R言語による統計的なデータ分析
  - R言語データ分析入門
  - R言語文法入門
  - R言語で代表的な統計的手法の実装

- C++マスター編
  - C++入門
  - C++クラス・オブジェクト指向入門
  - C++中級
  - C++徹底演習

主催講座一覧(2)

- ディープラーニング入門編
  - ディープラーニングのための数学
  - ディープラーニング理論入門
  - tensorflow入門
  - ディープラーニング実装入門(tensorflow)
  - CNN入門
  - 強化学習入門
  - RNN・LSTM入門

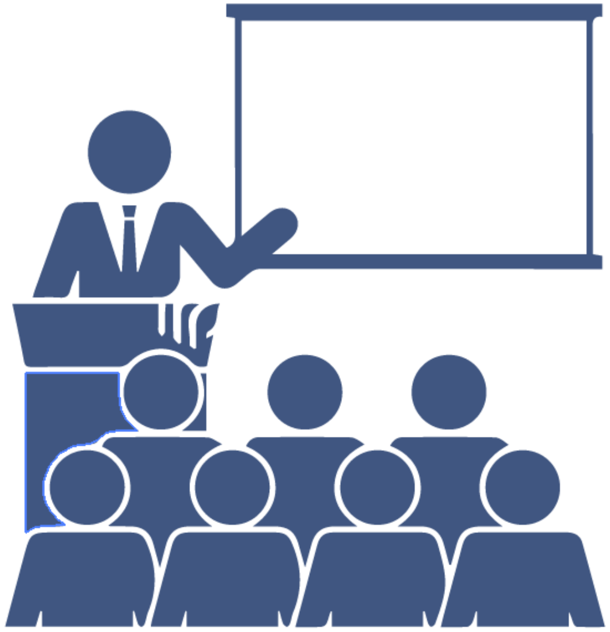
- 自然言語処理編
  - 前処理
  - 分散表現入門
  - wor2vec入門
  - トピックモデル入門

  セット講座

- イベント・勉強会
  - AIビジネス創出アイディアソン
  - Python/C++勉強会（毎週月曜）

- その他講座
  - Swiftによるiosアプリ開発
  - Unity入門
  - Unity実用
  - seq2seq
  - JavaScript初心者のためのWebフロントエンド入門

社員研修プログラムのご依頼について



➢ これから会社を牽引できるデータ人材の育成を行います。

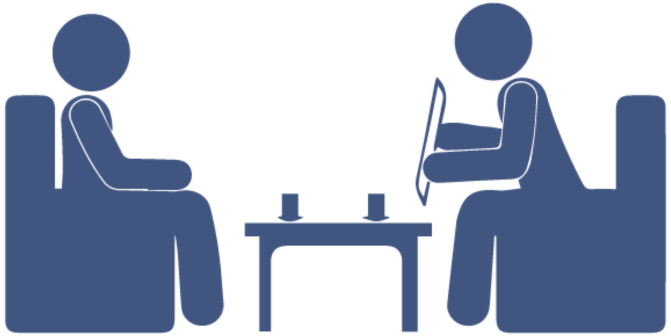
➢ 提供講座例

・Python基礎講座
・統計学/R言語講座
・機械学習/Deep Learning実践講座

➢ よくあるご質問

・少人数からでも大丈夫です。
・必要人材と会社の進路から相談させていただきます。
・1回から最大20回まで知識から実践までプログラムを用意させていただきます。
・費用等のご相談もお気軽にお聞きください。

人工知能/機械学習の導入および実装のご相談



➢ 会社で新事業を導入されたい方

➢ <u>こんな悩みをお持ちの方へ</u>

・新事業を始めたいけど、データに関する知識がない。
・データ人材が不足していて実装ができない。
・今後、どんな事業を展開してばいいかわからない。

➢ <u>提案サービス例</u>

・データ分析、プログラミング実装
・機械学習、人工知能を使った事業の相談
・費用、期間等のご相談もお気軽にお聞きください。

子供にプログラミングを習わせませんか

- 小中高生向けのプログラミングスクール・家庭教師はじめました！
- お子様のレベルに合わせたコースをお選び頂けます。

パソコン入門コース

Webクリエイターコース

ゲームクリエイターコース

エンジニアコース

アンケートのお願い

講座の改善のため、以下のURLからアンケートの協力をお願いしております

https://seminar.to-kei.net/qt/?pytime

仕事のご依頼・ご相談は
info@avilen.co.jp
までお問い合わせください