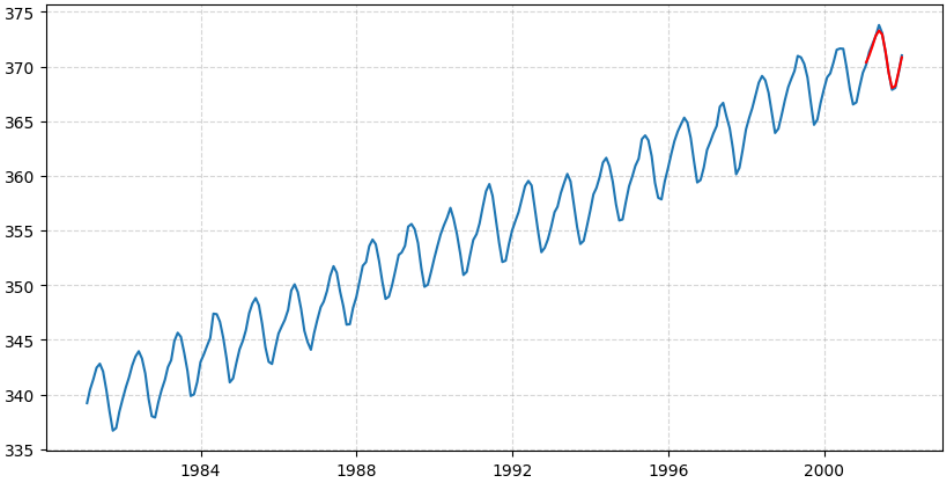
GCI 2024 Winter
Week8 時系列データとモデリング

松尾・岩澤研究室
Matsuo-Iwasawa Lab UTokyo

講師・スライド作成：石田将貴

今回の目標

今回の目標はCO_2濃度の予測をARIMAモデルを用いて行うことが目標です。



予測に必要な要素

- 時系列データの可視化
- STL分解
- 定常性
- 自己相関性
- 時系列モデリング(ARIMA)

高度な内容を含むので、講義内に理解するよりも、講義後にご自身で時系列データを扱う際に思い出せるように目指してください！

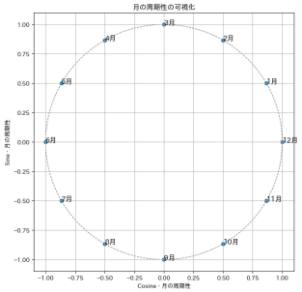
時系列データの処理

復習

モデルが入力に使いやすい形

	No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
25560	25561	2012	12	1	0	41.0	-16	-8.0	1035.0	NW	1.79	0	0
25561	25562	2012	12	1	1	46.0	-17	-7.0	1035.0	cv	0.89	0	0
25562	25563	2012	12	1	2	37.0	-15	-7.0	1036.0	cv	1.34	0	0
25563	25564	2012	12	1	3	48.0	-15	-9.0	1035.0	NE	0.89	0	0
25564	25565	2012	12	1	4	43.0	-14	-9.0	1035.0	NE	1.78	0	0

周期性を考慮した
エンコーディング
week7 4.3



特徴量を作ったり可視化に向く形
(時系列変化を追いやすい)

week7 4.1

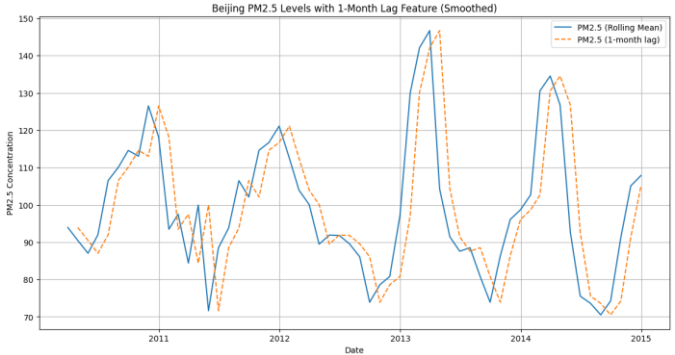
	pm2.5
2012-12-01 00:00:00	41.0
2012-12-01 01:00:00	46.0
2012-12-01 02:00:00	37.0
2012-12-01 03:00:00	48.0
2012-12-01 04:00:00	43.0

特徴量
生成

	pm2.5
2012-12-01	NaN
2012-12-02	2.630252
2012-12-03	0.127796
2012-12-04	0.975000
2012-12-05	0.743590

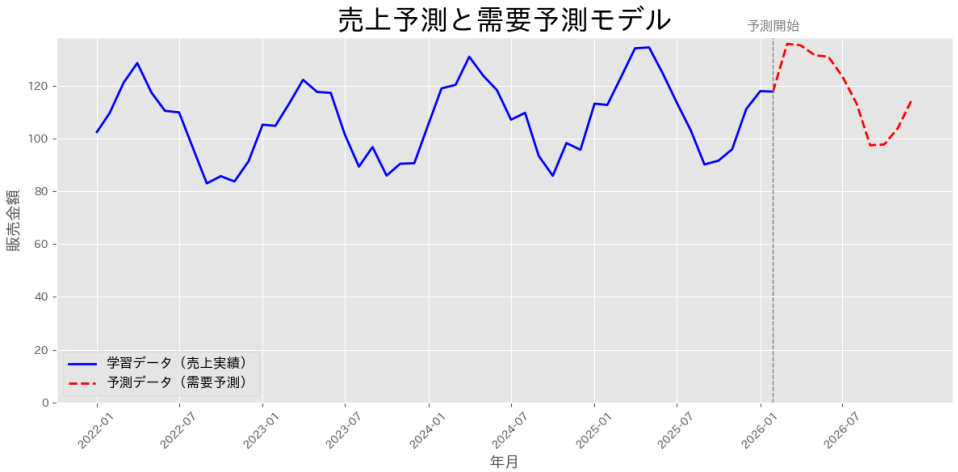
week7 4.2

可視化
(week5)

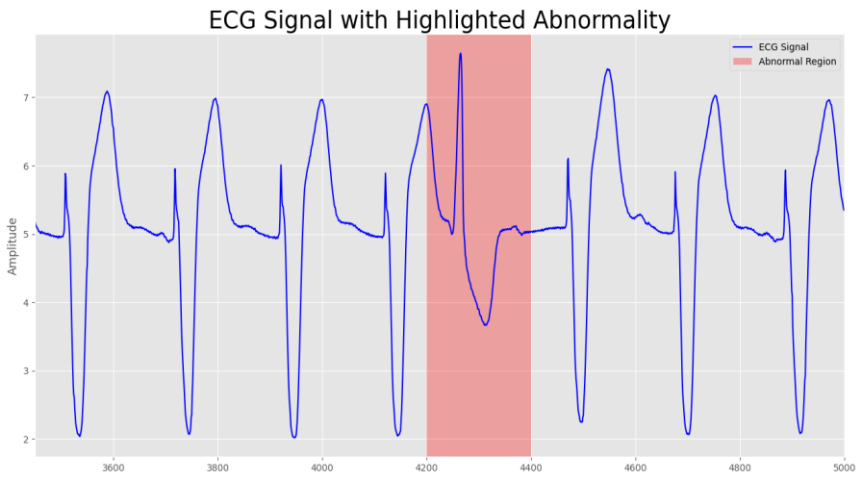


時系列解析のモチベーション

ビジネス応用例として以下のような目的があります。
①**目的変数の将来予測**　例：売上予測、需要予測、在庫管理の最適化
②**異常検知**　例：工場設備の故障検知、電力使用量の異常検出、心電図の異常心拍検出



将来予測



異常検知

時系列モデリングについての概観

統計モデル：時系列データが特定の統計的性質を持つとして予測
機械学習モデル：統計的性質を仮定せず、非線形な関係を捉えながら予測

	カテゴリ	手法	概要	特徴
	統計モデル系	ARIMA系（回帰モデル系）	自己回帰（AR）、差分化（I）、移動平均（MA）を組み合わせた時系列予測モデル	- 定常性が前提 - トレンドや季節性を除去可能 - 過去の値と誤差を活用
		SSM（状態空間モデル系）	状態変数（観測できない潜在変数）と観測データを動的にモデル化	- ノイズや動的変化に強い - カルマンフィルタなどで逐次推定可能
機械学習モデル系(例)	非時系列手法	ランダムフォレスト	複数の決定木を用いたアンサンブル学習	- 過学習に強い - 特徴の重要度を可視化可能 - 高い汎用性
	時系列手法	RNN	過去の情報を隠れ層に保持し、現在の出力を計算するニューラルネットワーク	- 短期依存をモデル化するのが得意 - 長期依存は苦手（勾配消失問題）
		Transformer	自己注意機構（Self-Attention）を用いて並列処理可能な深層学習モデル	- 長期依存も効率的に扱える - 自然言語処理（NLP）など広範な応用分野

統計モデルに必要な要素

統計モデルは「予測する時系列データが、過去の値に依存して将来の値が決まる」ことを前提にしています。

定常性

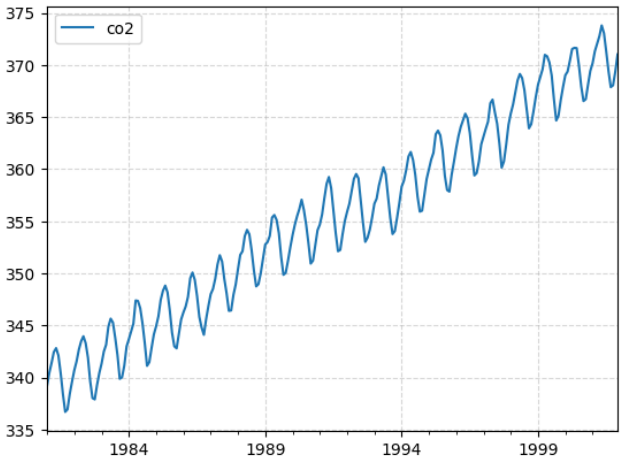
- **データの性質が時間によらず一定**
- **分散、平均値、自己共分散が一定**
- **非定常なデータは定常なデータに変換する処理が必要**

自己相関

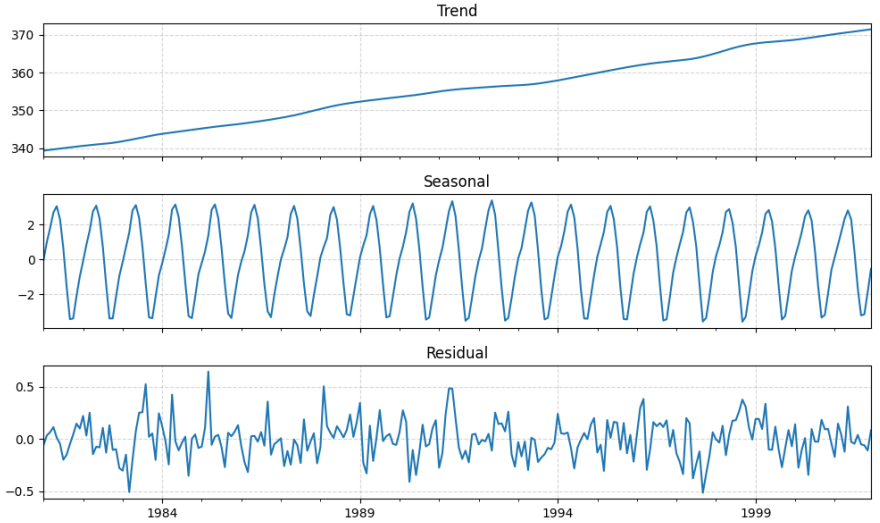
- **過去の観測値が将来の観測値に影響を与えている**
- **自己相関係数をプロットして確認**

時系列データの構成要素とSTL分解

時系列データはトレンド成分、季節成分、残差成分から構成されていることが多いです。それぞれの成分に分ける一つの方法がSTL分解です。



STL分解



notebookへ

統計モデルに必要な要素

統計モデルは「予測する時系列データが、過去の値に依存して将来の値が決まる」ことを前提にしています。

定常性

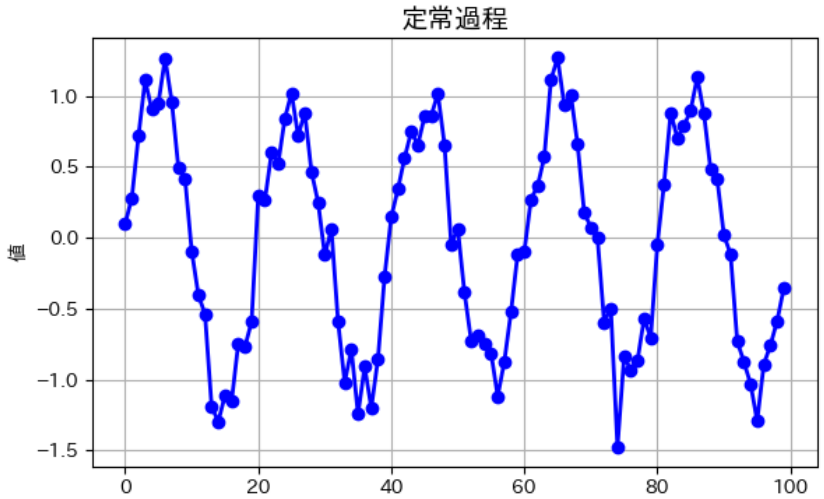
- **データの性質が時間によらず一定**
- **分散、平均値、自己共分散が一定**
- **非定常なデータは定常なデータに変換する処理が必要**

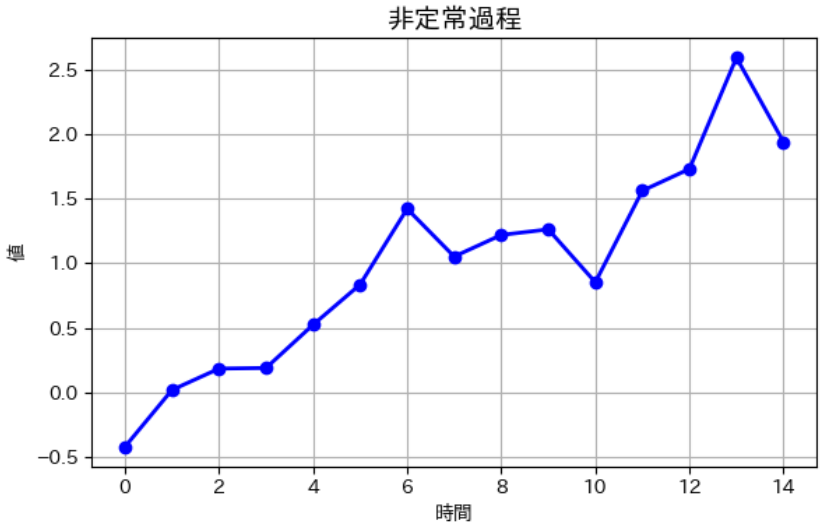
自己相関

- **過去の観測値が将来の観測値に影響を与えている**
- **自己相関係数をプロットして確認**

定常・非定常

(弱)定常性を満たすデータは、分散、平均値、自己共分散 (ラグのみに依存) が一定である性質を持ちます。



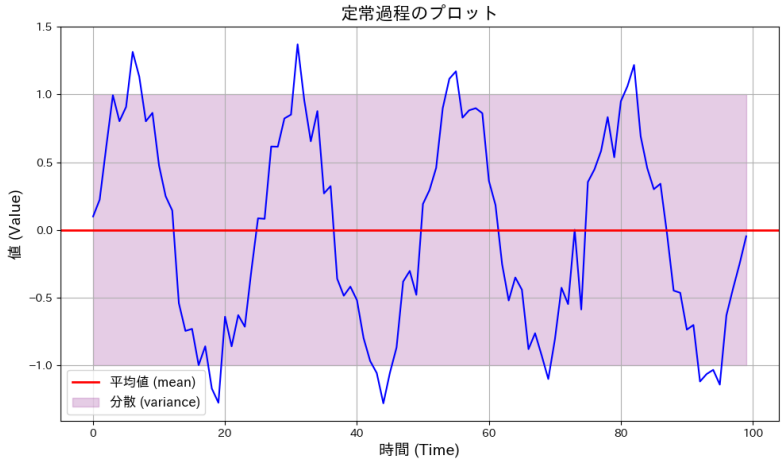


各時点での時系列データの平均値が一定であり、自己共分散が時間差にのみ依存する

左の性質が当てはまらないデータが非定常なデータ

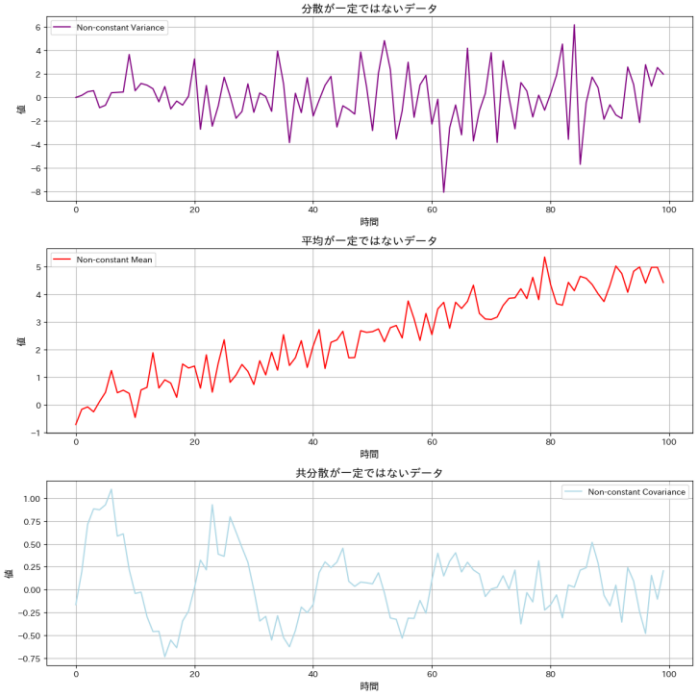
定常・非定常

(弱)定常性を満たすデータは、分散、平均値、自己共分散が一定(ラグのみに依存)である性質を持ちます。



分散が一定

平均値が一定



ラグにのみ依存

各時点での時系列データの平均値が一定であり、
自己共分散が時間差にのみ依存する

統計モデルに必要な要素

統計モデルは「予測する時系列データが、過去の値に依存して将来の値が決まる」ことを前提にしています。

定常性

- データの性質が時間によらず一定
- 分散、平均値、自己共分散が一定
- 非定常なデータは定常なデータに変換する処理が必要

自己相関

- **過去の観測値が将来の観測値に影響を与えている**
- **自己相関係数をプロットして確認**

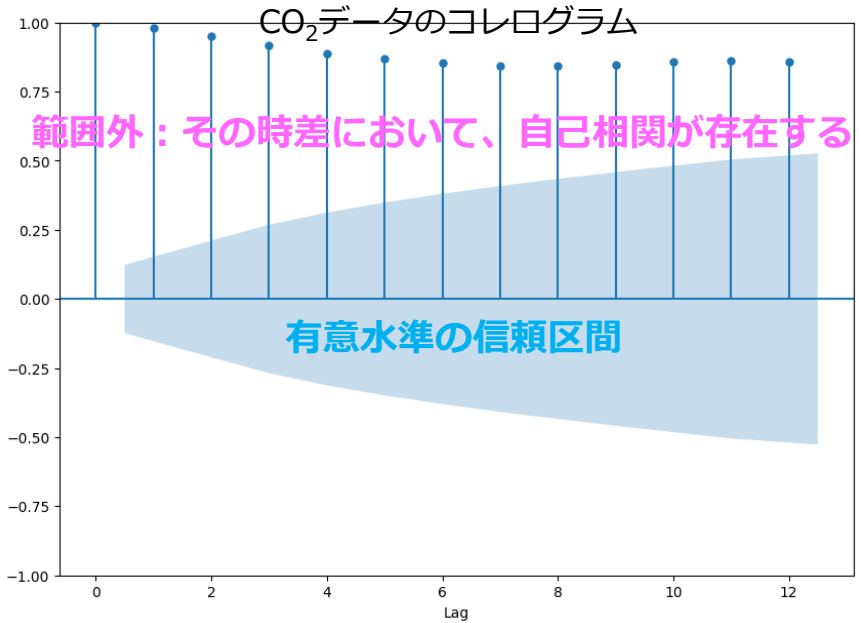
ラグのみに依存するとは？自己相関について

自己相関係数：現在のデータとk期前のデータとの相関を計算
データにトレンドや季節性がある場合、それらが自己相関に反映されます。

$$\rho_k = \frac{\sum_{t=k+1}^{n}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

自己相関係数を計算することで、
過去のデータが現在のデータに
どの程度影響しているかを把握
することができる

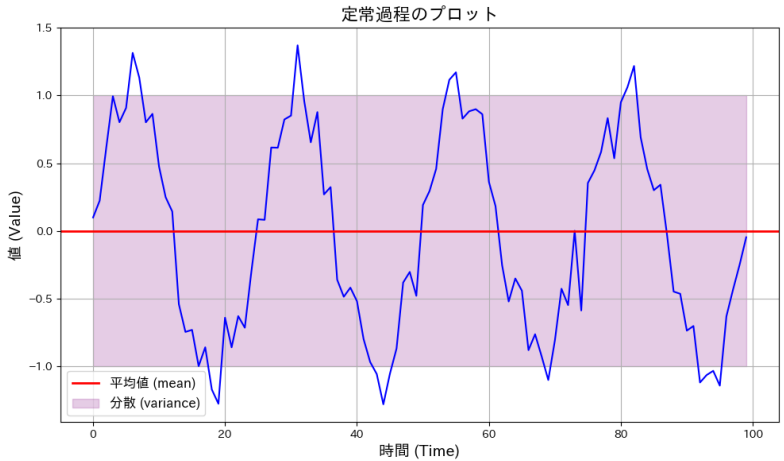
```
#自己相関を計算・プロットするplot_acf関数をインポート
from statsmodels.graphics.tsaplots import plot_acf

fig, ax = plt.subplots(figsize=(10, 7))
#自己相関をプロット。lags=12と指定することによって、1年分(1周期分)の自己相関を計算
plot_acf(co2_data, lags=12, ax = ax)
plt.xlabel('Lag')
plt.show()
```



自己相関係数が時間とともに緩やかに減衰せず、全体的に高い値を持っている
→この時系列データはトレンドを持つ可能性が高い

定常性の確認- ADF検定

弱定常性かどうかを検証する仮説検定がADF検定です。
単位根を持つ時系列データかどうかを検証します。



ラグにのみ依存

ADF検定

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \ldots + \delta_p \Delta y_{t-p} + \epsilon_t$$

分散が一定

γ=0 であれば単位根があり、
データは非定常とする

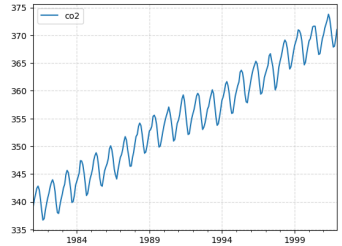
平均値が一定

※単位根を持つ時系列データ(過去の値に大きく依存し、
時間が進むにつれて平均や分散が一定ではなくなります。

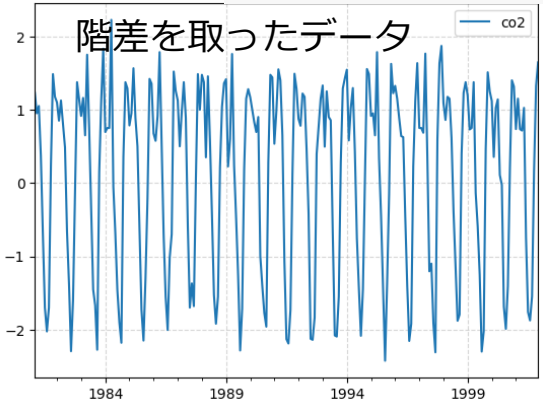
定常なデータに変換する

階差を取ることで定常なデータに変換することができる場合もあります。

元データ

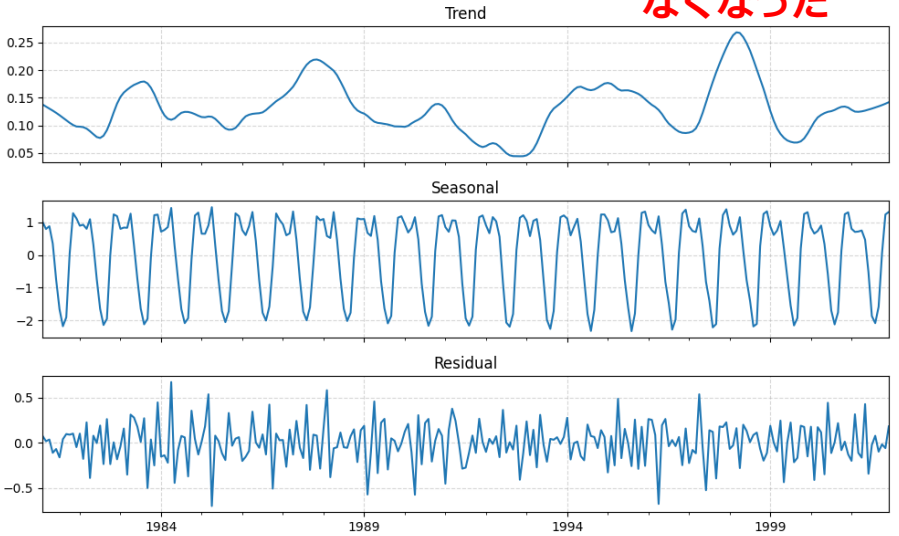



```
#隣り合う時点の差のデータを作成
co2_data_diff = co2_data.diff().dropna()
```



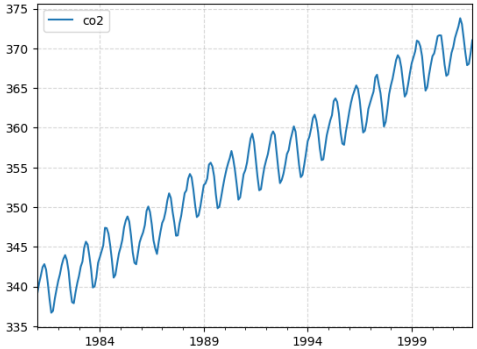
STL分解

トレンド成分がなくなった

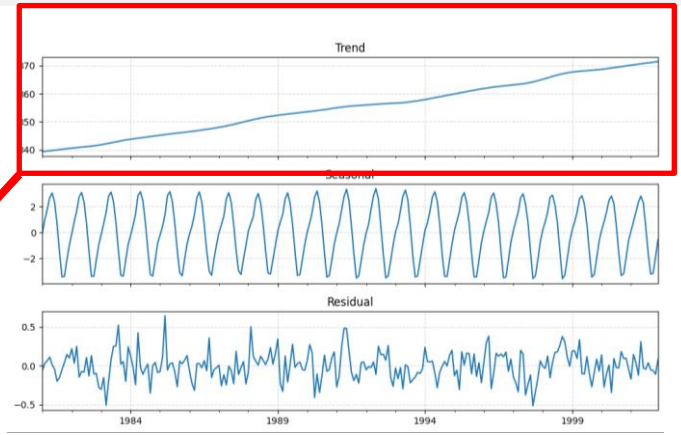


差分を取って定常にする

階差を取ることでトレンドや季節成分を除去することができるケースがあります。



STL分解



$$y_t = 5t + \epsilon_t$$

トレンド成分(時間とともに増加)

誤差項

$$y'_t = y_t - y_{t-1} = (5t + \epsilon_t) - [5(t-1) + \epsilon_{t-1}]$$

$$y'_t = 5 + (\epsilon_t - \epsilon_{t-1})$$

平均値が一定になる

notebookへ

ARIMA

自己回帰（AR）と移動平均（MA）を組み合わせたARMAモデルにおいて、
階差を取り非定常なデータでも扱えるようにした(I)ものがARIMAモデルです。

ARIMA

ARIMA

ARIMAモデルでは3つのパラメタを指定する必要があります。
パラメータの最適値はグリッドサーチなどを用いて探索します。

ARIMA (p, d ,q)

※グリッドサーチに関してはweek9で学んでください

AR(p):過去のデータ（ラグ）を何回分参照するか　（自己回帰）
I(d):トレンドや非定常性を取り除くために差分を取る回数　(差分)
MA(q):過去の予測誤差（残差）を何回分参照するか　（移動平均）

例: ARIMA(1, 1, 1)
p=1: 1つ前の値を利用
d=1: 1回差分をとる
q=1: 1つ前の誤差を利用

$$y'_t = \phi_1 y'_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}$$

一次差分

ϕ：自己回帰係数
θ：移動平均係数

AIC

パラメタの選定に今回はAICを用います。AICはモデルの複雑さと適合度をトレードオフとして評価する指標になります。

尤度に関してはweek5 統計・確率の2.11.4 などを参照.
モデルのパラメタが与えられたときに、観測データが得られる確率

$$AIC = 2k - 2\ln(L)$$

モデルのパラメタ数(モデルの複雑さ)

尤度(モデルがデータをどれだけよく説明しているか)

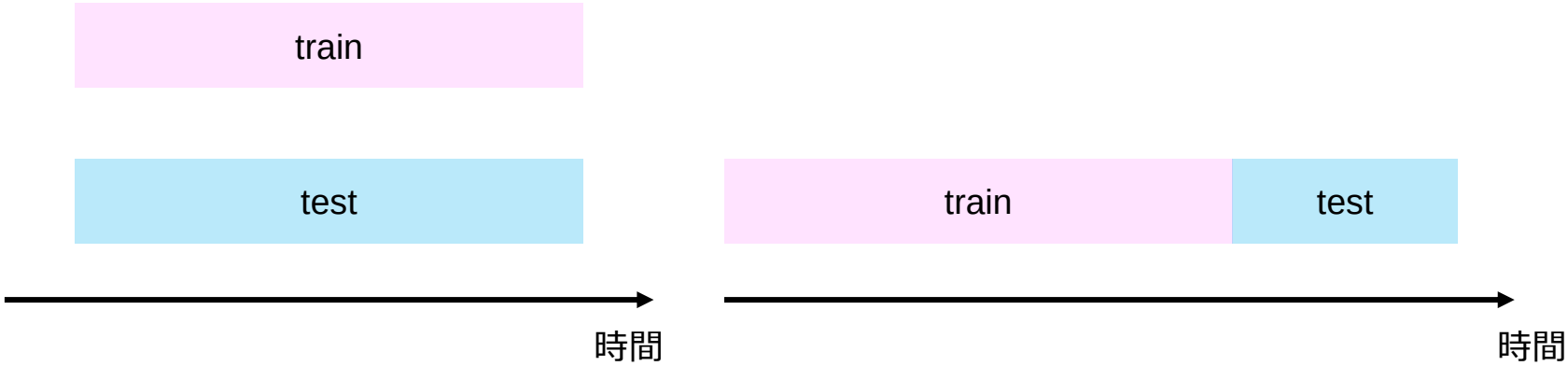
モデル	p	d	q	k	−2ln(L)	AIC
ARIMA(1, 1, 1)	1	1	1	3	100	100+2×3=106
ARIMA(2, 1, 2)	2	1	2	5	90	90+2×5=100
ARIMA(3, 1, 3)	3	1	3	7	88	88+2×7=102

AICを最小化することで、過学習を防ぎつつ適合度が高いモデルを選択できる

時系列情報の扱い方

復習

基本的に時系列データを用いる場合、未来予測の場合が多い(右図)
その場合、学習データに未来のデータが入らないように気を付ける必要がある(リークになる)



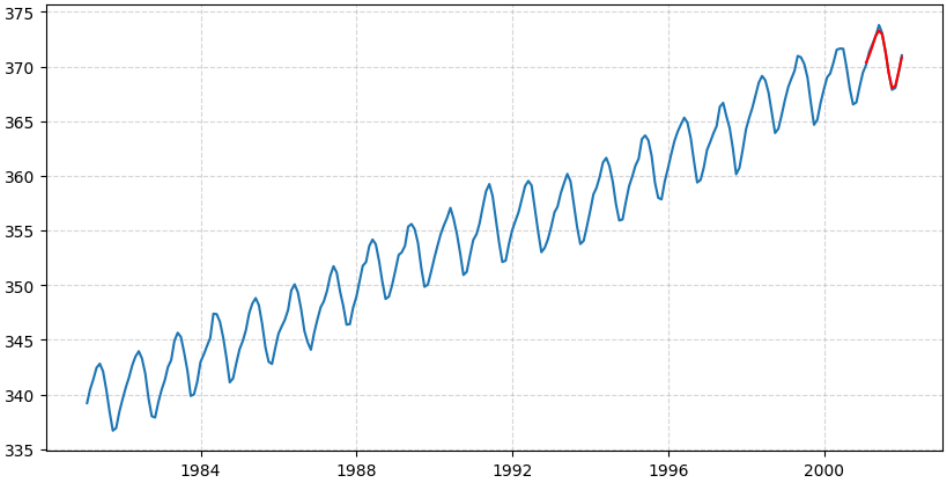

```
# 訓練データとテストデータを準備
len_test = 12
train = co2_data[:len(co2_data) -len_test]
test = co2_data[len(co2_data)-len_test:]
```

時系列順に並んでいるデータ
最新の1年分のデータをテストデータにする処理を行う

notebookへ

今回の目標

今回の目標はCO_2濃度の予測をARIMAモデルを用いて行うことが目標です。



予測に必要な要素

- 時系列データの可視化
- STL分解
- 定常性
- 自己相関性
- 時系列モデリング(ARIMA)

高度な内容を含むので、講義内に理解するよりも、講義後にご自身で時系列データを扱う際に思い出せるように目指してください！

更なる学習のために

今回の講義は、時系列解析の中では基礎にあたる部分になります。
この教材では足りない部分も多くあるので、いくつか項目を挙げます。

モデルの精度を比較する

- **AICはモデルの妥当性を評価するが予測性能を直接評価しているわけではないのでMSEなどで予測誤差を図ることも必要**
- **単体モデルではなく、複数モデルやベンチマークでの評価も重要**

他のモデル

- **今回のデータの場合、季節成分も考慮するSARIMAモデルがより有効である可能性**
- **SSM（状態空間モデル）**
- **RNNやLSTM,Transformerなどの深層学習モデルなど**

前処理その他

- **自己相関と共に偏自己相関も確認する**
- **フーリエ変換・時間窓などの処理**
- **グレンジャー因果検定などによる変数同士の関係の探索**

更なる学習のために- 参考書例(講師の独断です)

- Pythonによる時系列分析: 予測モデル構築と企業事例
  髙橋 威知郎 (著)

- Pythonによる時系列予測
  (Compass Data Science)
  Marco Peixeiro (著)

- Pythonによる異常検知
  曽我部 東馬 (著), 曽我部 完 (監修)