



時系列分析入門

株式会社データミックス
2019年1月 ver

講義のゴール

- なぜ回帰分析では駄目なのか？
- 時系列分析の以下のトピックを理解し、専門書を読めるようにすること
 - ARモデル
 - MAモデル
 - ARMA, ARIMAモデル
 - SARIMAモデル
 - VARモデル
 - 機械学習を使ったモデル

アジェンダ

Day1

- 第1章 なぜ回帰分析ではいけないのか？
- 第2章 時系列分析の概要
- 第3章 定常性と自己相関
- 第4章 ARモデルとMAモデル
- 第5章 トレンドと単位根過程

Day2

- 第6章 季節調整済時系列モデル
- 実践課題
- 第7章 VARモデル

Day3

- 第8章 機械学習を使った時系列データ予測
- 総合演習



第1章

なぜ回帰分析ではいけないのか？

回帰分析をやってみてください

- data/bikesharing_hour.csv を読み込んでください。
- このデータはバイクのsharingサービスの利用者数の時間別データです。
- 被説明変数をcasual(未登録ユーザー)の利用人数とした時に、回帰分析をしてください。

データの説明

- instant: record index
- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from [\[Web Link\]](#))
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- + weathersit :
 - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
 - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
 - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
 - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius.
- atemp: Normalized feeling temperature in Celsius.
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)

<target>

- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

出処: <http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

回帰分析の仮定

回帰分析はシンプルな分析ですが、実は正しく使うためには「回帰分析の仮定」をチェックする必要があります。もし仮定を満たさない場合は推定結果が間違っています。

回帰分析が仮定している条件とそのチェック方法を紹介します。

線形回帰モデルが想定する仮定

$$Y = \alpha + \beta x + \varepsilon$$

$$\varepsilon \sim N(0, \sigma)$$

- 被説明変数は説明変数と傾きの掛け算を足したもので説明できる
- 誤差（ノイズ）が正規分布に従う
- 誤差（ノイズ）が説明変数に関わらず等分散
- 誤差（ノイズ）に系列相関がない

ちなみに残差と誤差の違いは

- ・「誤差」は確率変数
- ・「残差」は誤差の推定値

$$Y = \alpha + \beta \times X + \varepsilon$$

被説明変数 切片 回帰
係数 説明
変数 誤差

推定

残差は
誤差の推定値

$$Y = 2.0 + 3.5 \times X + \hat{\varepsilon}$$

切片の
推定値

係数の
推定値

残差

誤差の正規性のチェック

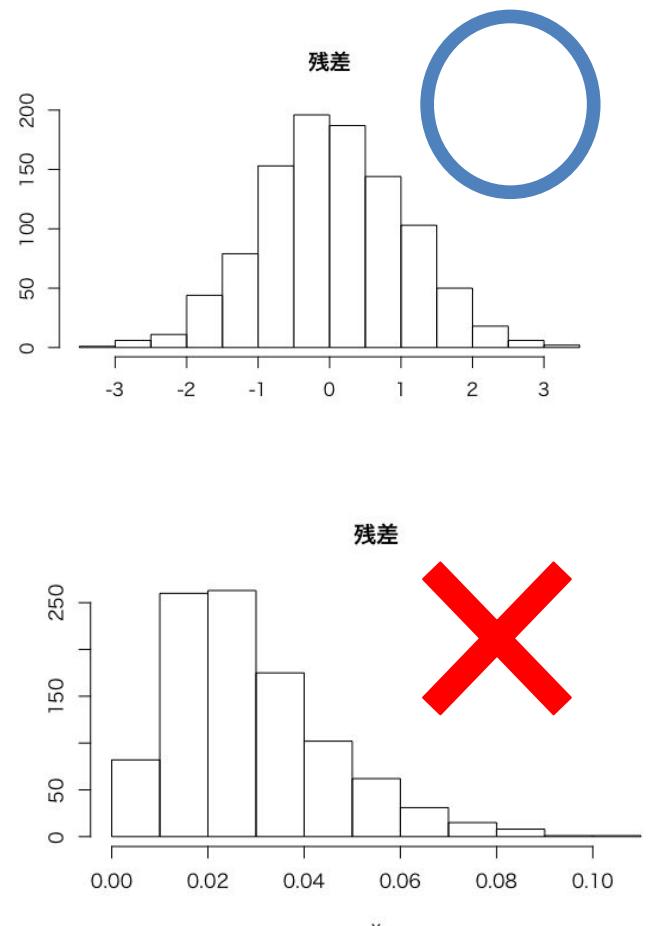
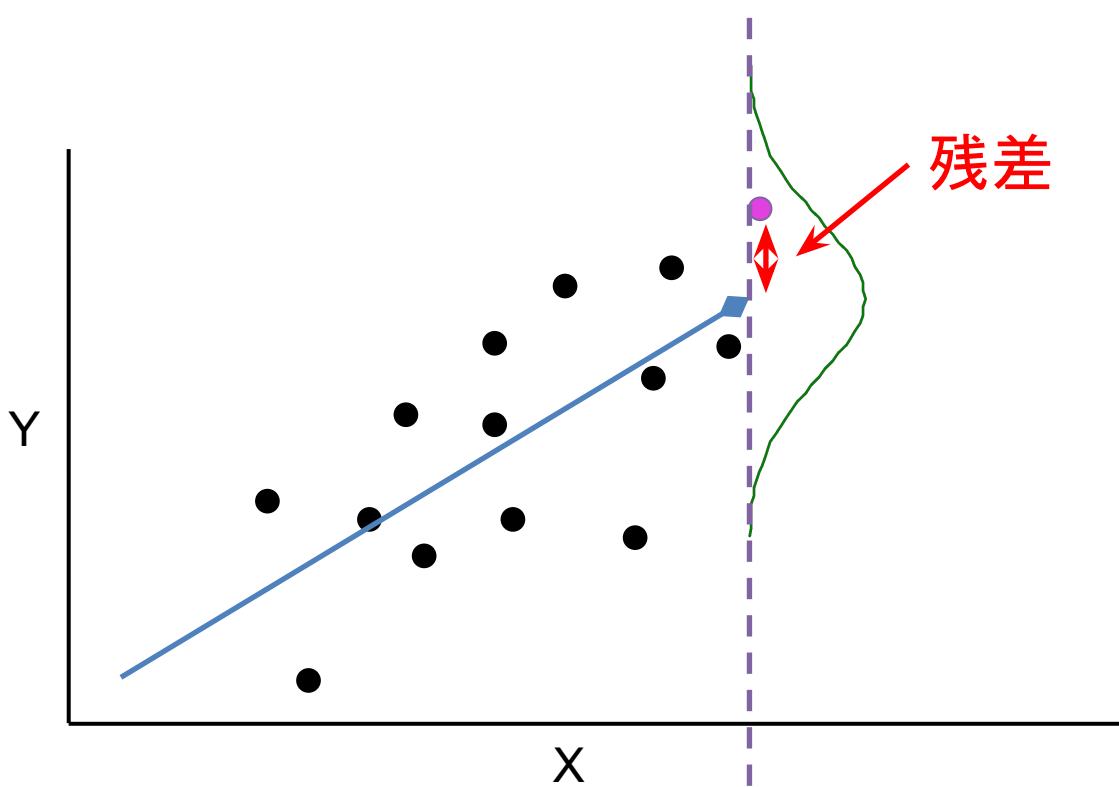
$$Y = \alpha + \beta x + \varepsilon$$

$$\varepsilon \sim N(0, \sigma)$$

- 被説明変数は説明変数と傾きの掛け算を足したもので説明できる
- **残差誤差(ノイズ)が正規分布に従う**
- 誤差(ノイズ)が説明変数に関わらず等分散
- 誤差(ノイズ)に系列相関がない

残差が平均0の正規分布に従うとは

説明変数をある値で切った時に残差があるわけですが、残差だけを抜き取ってヒストグラムを作った時に正規分布になるか？



正規性をチェックする方法

グラフでチェックする場合

- Q-Q plot

検定したい場合(検定の名前は覚えなくて良いです。そんなこと言ってたなあくらいで良いです。ちょっと微妙なので……)

- Shapiro-Wilk Test(シャピオ-ウィルク検定)
- Kolmogorov-Suminov Test(コルモゴロフ-スミノルフ検定)
- Jarque-Bera Test(ジャック-ベラ検定)

例: Jarque-Bera検定

<Jarque-Bera検定>

帰無仮説: 対象のデータは正規分布

対立仮説: 対象のデータは正規分布ではない

Rなら簡単にJarque-Bera検定ができる

```
> library(tseries)  
> jarque.bera.test ( data = データ名 )
```

p-valueが0.05より大きければ、帰無仮説を棄却できない。正規分布ではないとは言い切れないで…（まあよしと正規分布に従うとするか…）みたいな話です

もし正規性の仮定を満たさないときは

- まずチェックすべきは**左右対称**か？
 - 左右対称なら推定値自体はおかしくない（統計学の言葉で不偏）
 - 左右非対称はよろしくない
 - 回帰係数自体が正しくない可能性がある
- いくつかの対処方法
 - 被説明変数を対数変換 :log()
 - 説明変数を加える
 - 外れ値を削除する

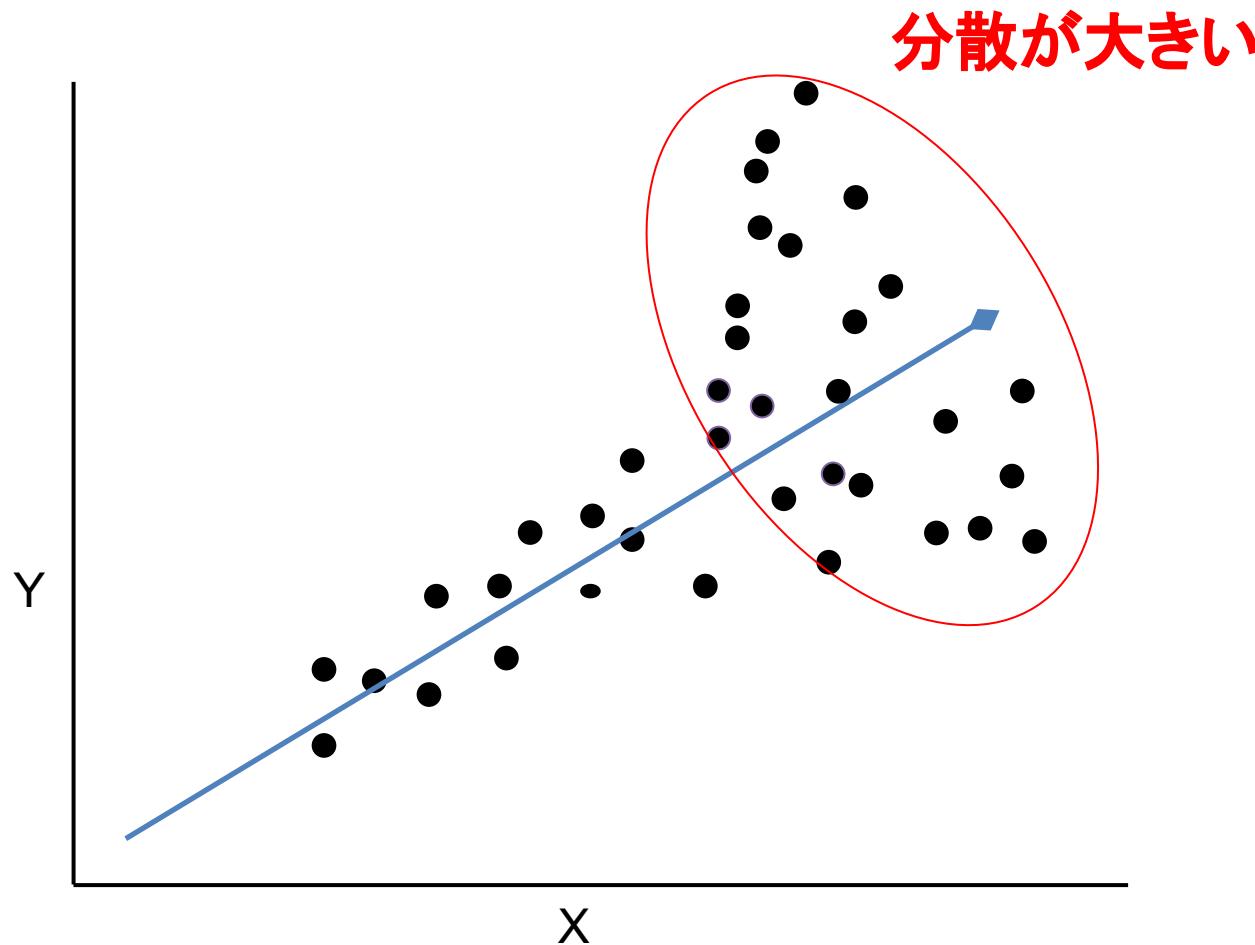
続いて、等分散性のはなし

$$Y = \alpha + \beta x + \varepsilon$$

$$\varepsilon \sim N(0, \sigma)$$

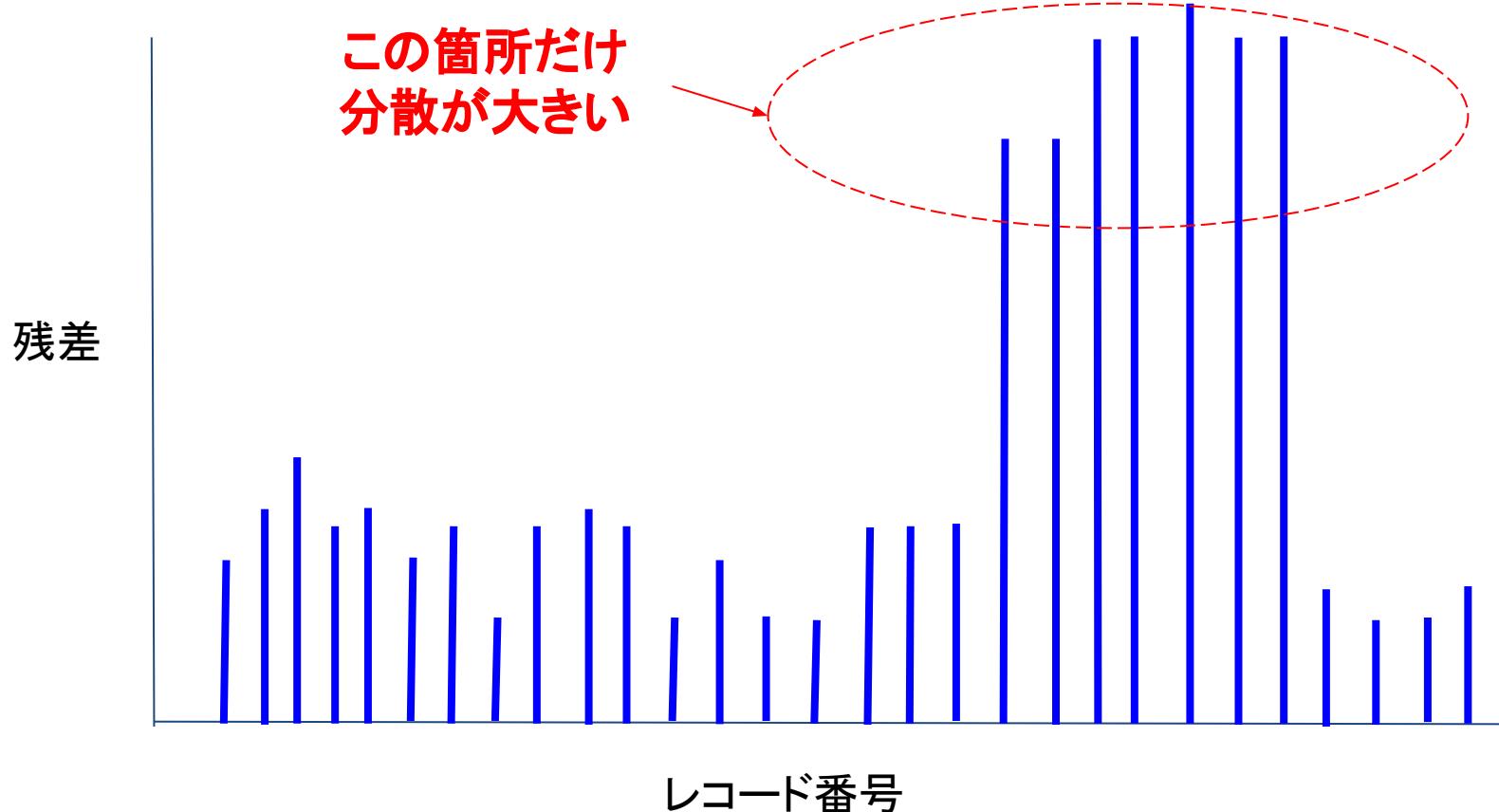
- 被説明変数は説明変数と傾きの掛け算を足したもので説明できる
- 残差誤差(ノイズ)が正規分布に従う
- 残差誤差(ノイズ)が説明変数に関わらず等分散
- 誤差(ノイズ)に系列相関がない

残差が等分散ではない状態とは



グラフでチェックする方法

横軸にレコード番号、縦に残差を取り可視化します。



Breusch-Pagan検定

<Breusch-Pagan検定>

帰無仮説：分散が均一である

対立仮説：分散が均一でない

Rなら簡単にBreusch-Pagan検定ができる

```
library(lmtest)  
>bptest(lmオブジェクト)
```

p-valueが0.05より大きければ、帰無仮説を棄却できないので、まあ分散が均一かもしれないね……という話になる

もし分散が均一でない場合は

- 推定値は大丈夫(不偏)
- ただし残差の分散がおかしいので、回帰係数に
関わるt検定やF検定はおかしい可能性がある。
- ではどうすれば?
 - 被説明変数を対数変換してみましょう(少しはマシにな
るかも)
 - Whiteの修正という高度な手法がある

最後に系列相関のはなし

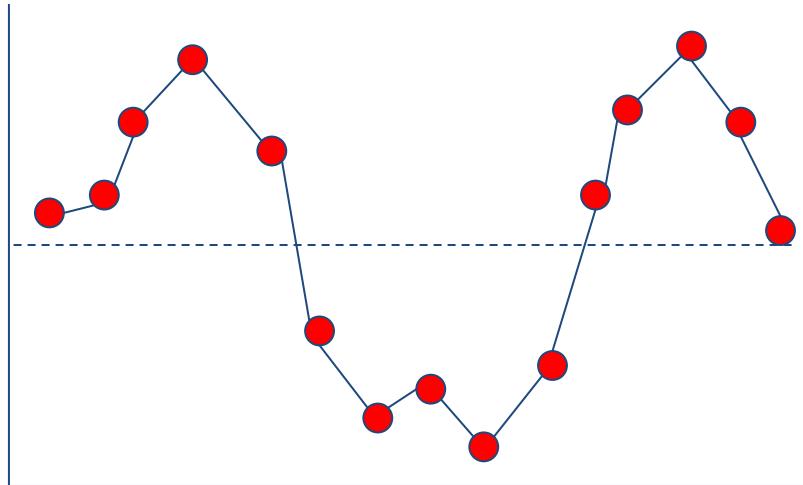
$$Y = \alpha + \beta x + \varepsilon$$

$$\varepsilon \sim N(0, \sigma)$$

- 被説明変数は説明変数と傾きの掛け算を足したもので説明できる
- 残差誤差(ノイズ)が正規分布に従う
- 残差誤差(ノイズ)が説明変数に関わらず等分散
- 残差誤差(ノイズ)に系列相関がない

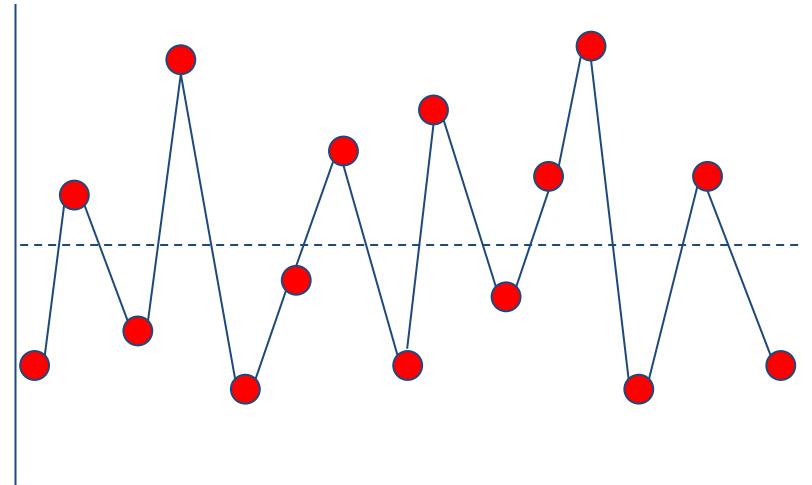
そもそも系列相関とは

系列相関とは残差どうしで相関があること。
残差をプロットした場合以下のようなグラフになると危ない



正の相関がある

ある程度同じ符号のまま推移する期間が続く

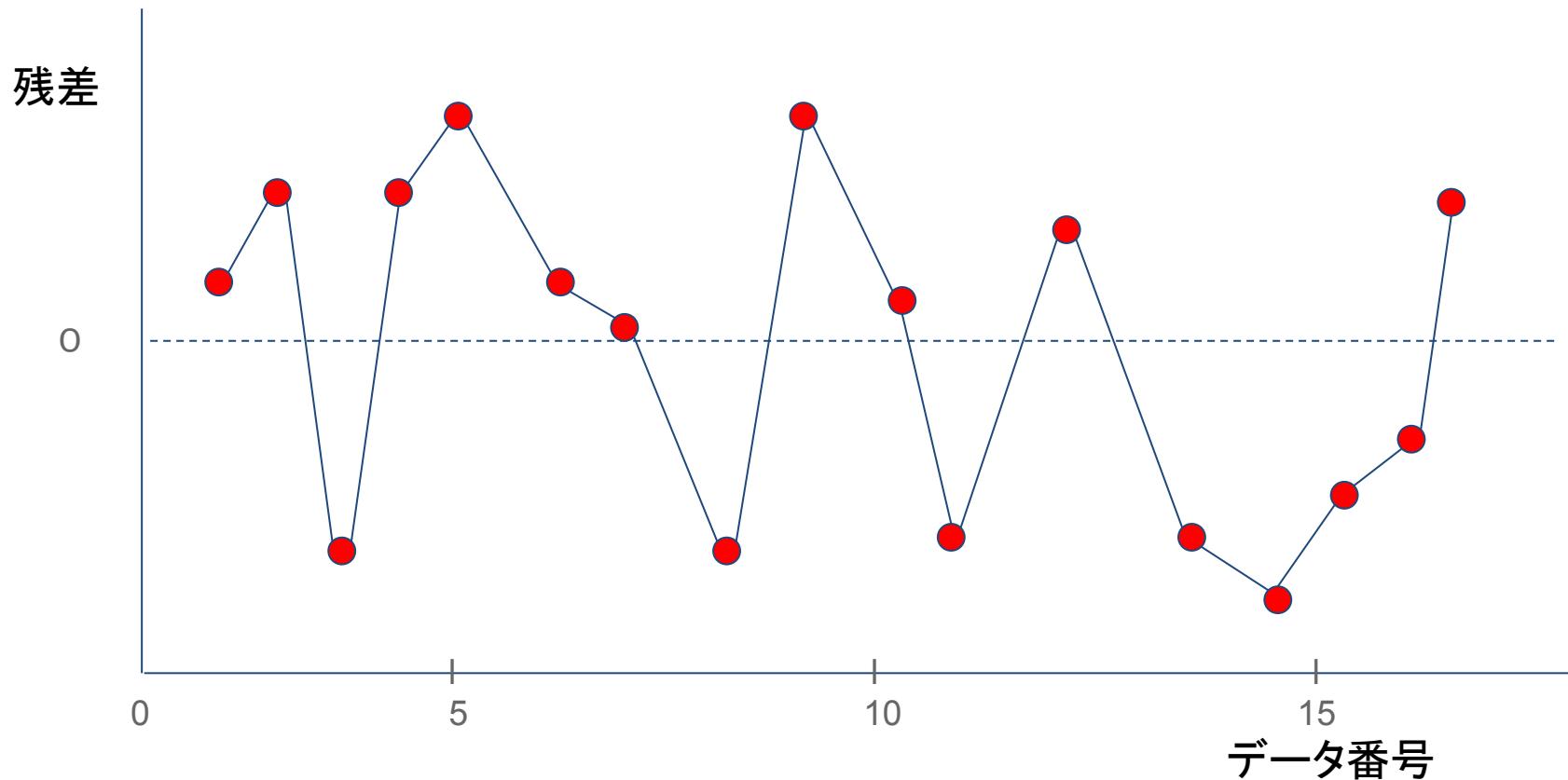


負の相関がある

ある程度符号が正と負を交互に推移する

グラフでチェックする場合

まずはグラフでチェック



Breusch-Godfrey検定

<Breusch-Godfrey検定>

帰無仮説：系列相関がない

対立仮説：系列相関がある

Rでは簡単にBreusch-Godfrey検定ができる

```
>bptest(lmオブジェクト, order =1)
```

p-valueが0.05より大きければ、帰無仮説が棄却できないので、まあ系列相関がないとは言えないかもね。

系列相関の仮定を破った場合

- 嬉しいことに回帰係数の推定値は間違っていない（不偏）
- しかしt検定の結果はめちゃくちゃ
 - 正の系列相関がある場合は、t値が上昇し棄却しなくて良い変数を棄却してしまいやすくなる
 - 負の相関がある場合はその逆
- ではどう対処するか？
 - 季節性を考慮するためのダミー変数を入れる

演習: 回帰診断

さきほど作った回帰モデルで

- 残差の正規分布か
- 分散が均一か
- 系列相関がないか

を確認してください。



第2章

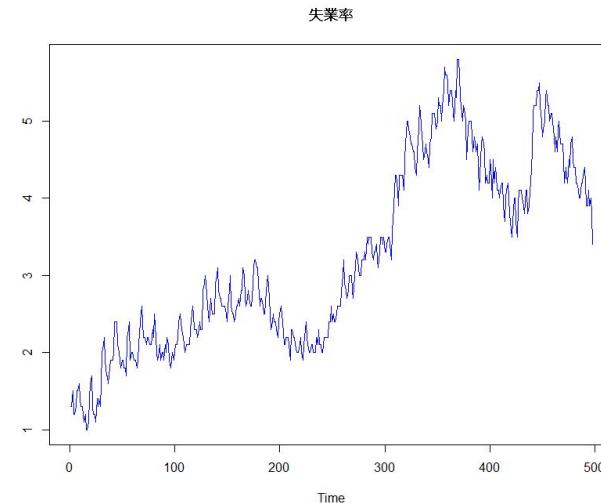
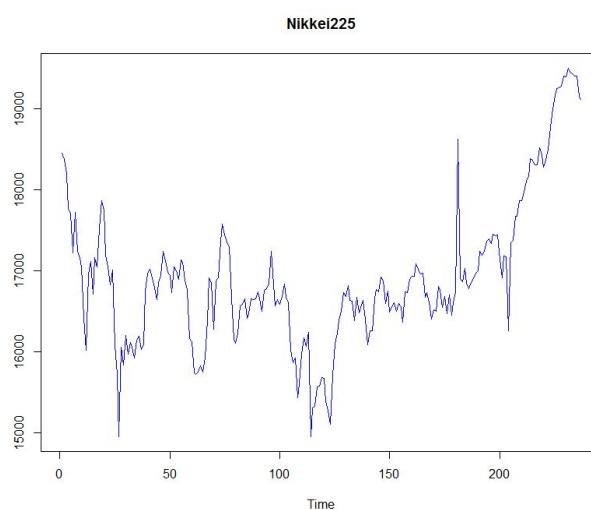
時系列分析の概要

時系列データとは？

- ・時間の推移と共に観測されるデータ
- ・観測される順序に意味がある

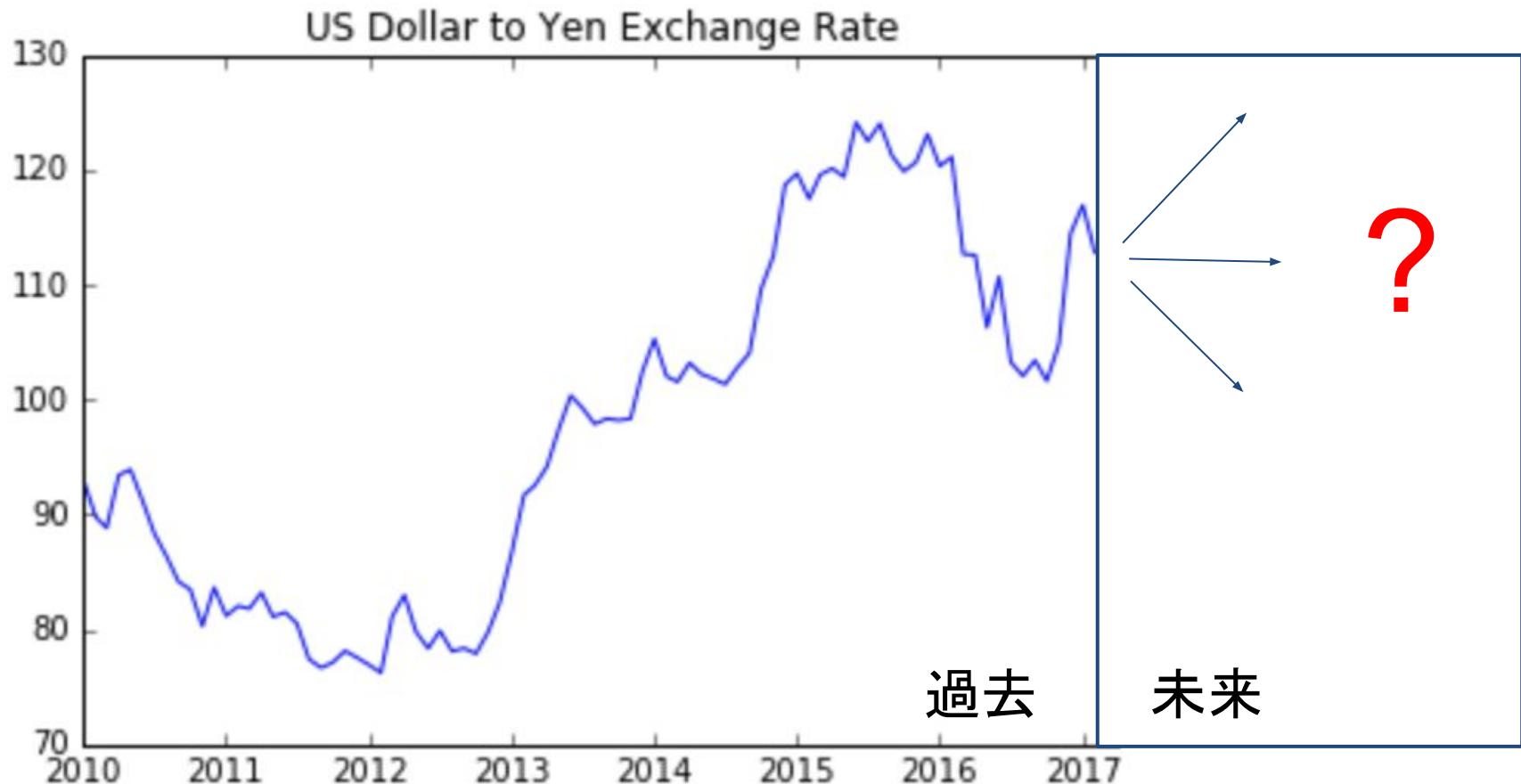
時系列データの例

株価、為替レート、GDP、失業率

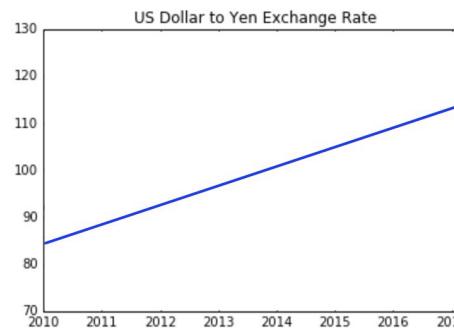
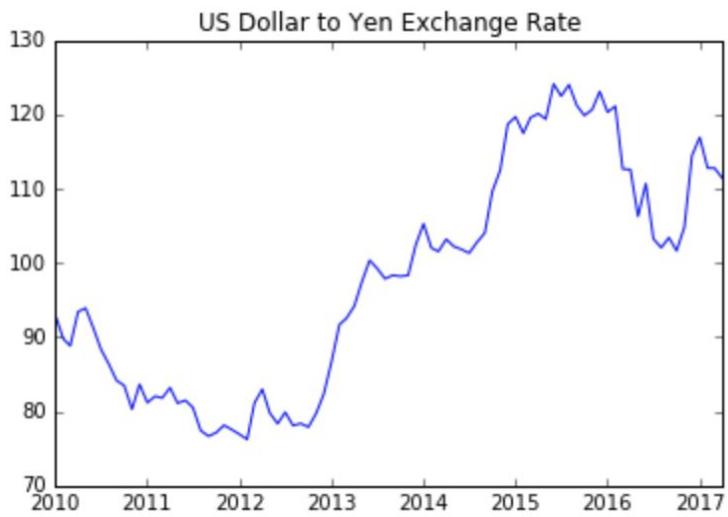


突然ですが、
未来を予測できたらどうしますか？

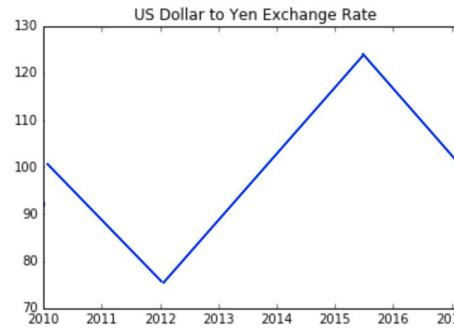
例えば、為替レートを予測できたら…



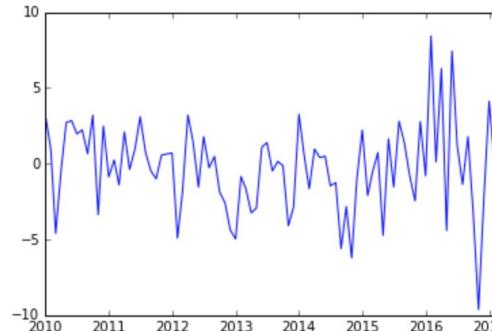
グラフを分解してみる



上昇してるなあ



周期性があるなあ



...(^_^;) ?

変動の要因

例えば、日経平均株価推移を回帰分析することを考える。回帰分析するに当たって株価変動に影響がありそうな要因をあげて見ると…

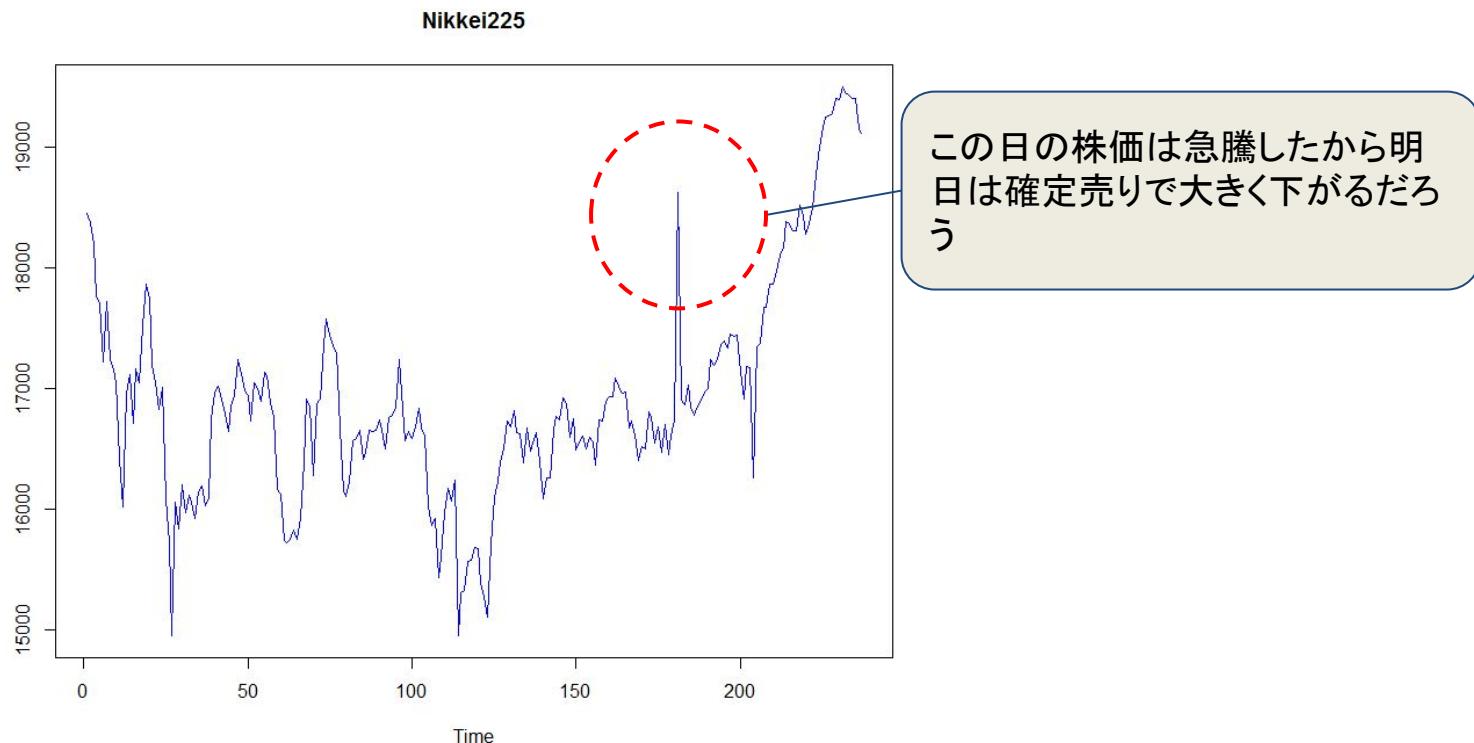
日経平均変動要因の例

前日の米ダウ株価、米ドル円、景気指数、長期金利、雇用者数・失業率、各構成銘柄のニュース、…

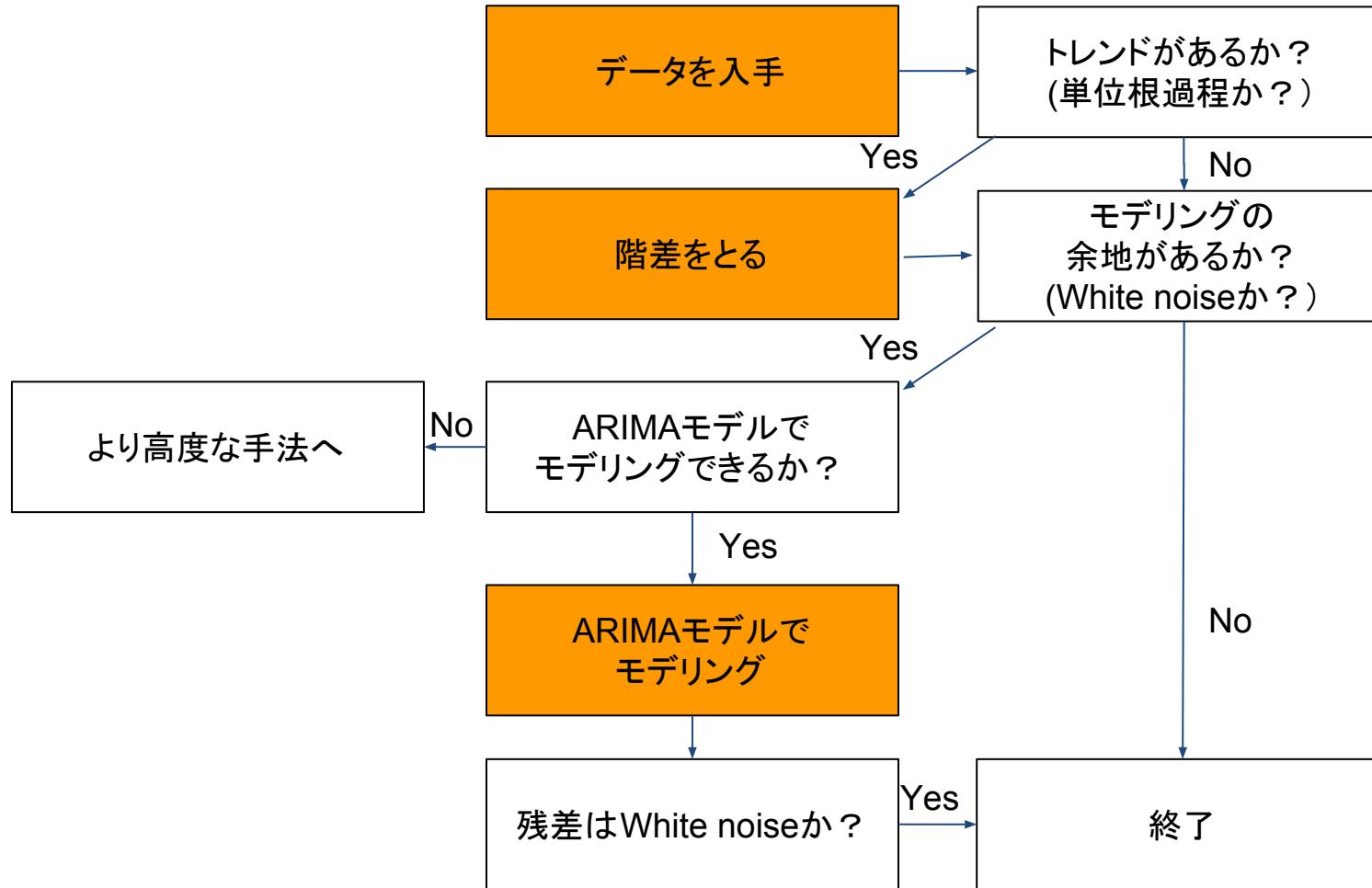
全ての要因は考慮できない

時系列分析の特徴

観測された値は多種多様な**要因が反映されてたもの**と考えられるため要因を考慮する必要がない。つまり、時系列分析では**データ自体が持つ構造を捉える**ことで予測を行う。



時系列分析の手順

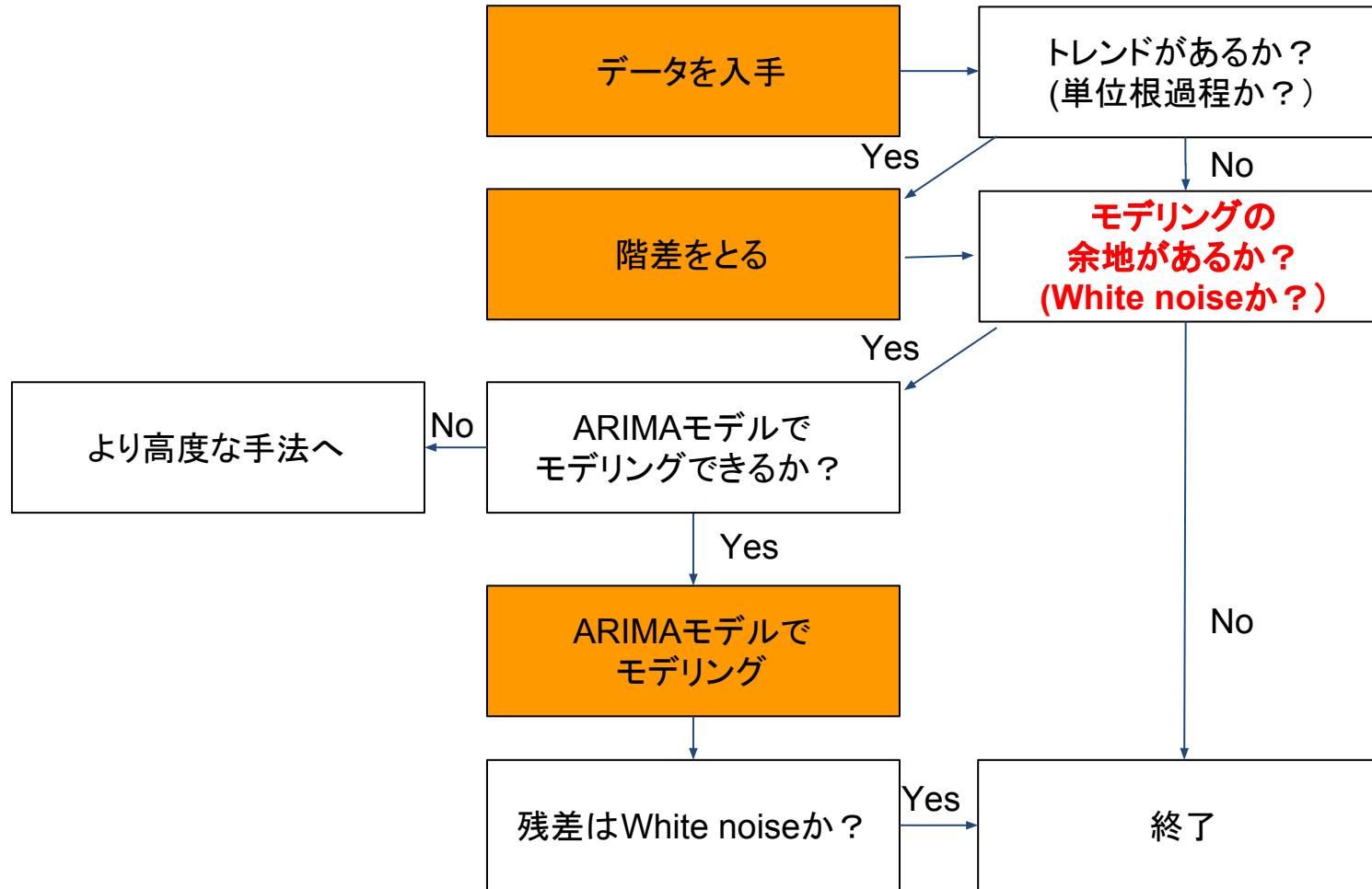




第3章

定常性と自己相関

時系列分析の手順



White noise : Rでやってみましょう1

White noiseのデータを作つてみましょう。

- サンプル数は200

<White noiseの式>

$$Z_{\{t\}} = \varepsilon \quad \varepsilon \sim N(0, \sigma)$$

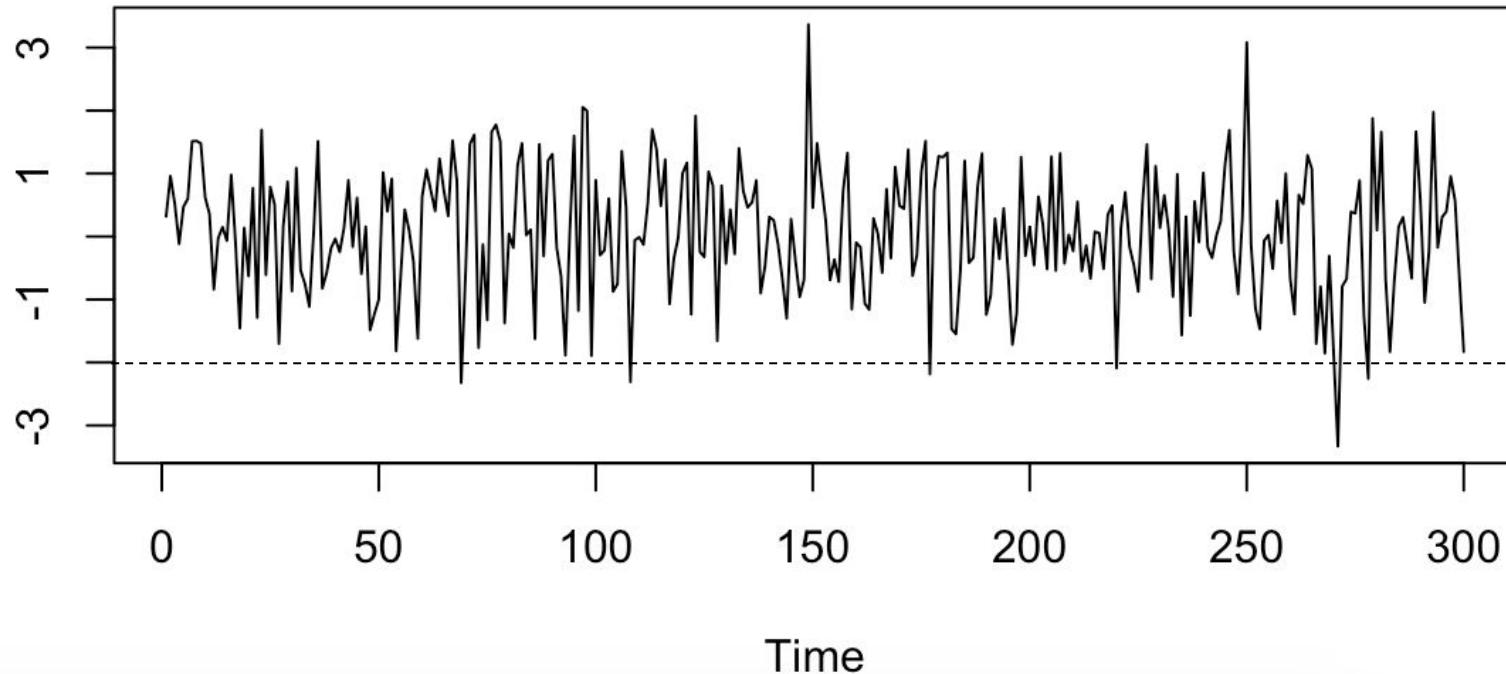
弱定常性はとても重要な概念

弱定常性とは

- 期待値は時間に依存しない
 $E(y_t) = \mu$
- 自己共分散は時間に依存せず、時間差のみに依存する
 $Cov(y_t, y_{t-k}) = \gamma_k$

定常性を日本語で言うと

定常過程: 時間が経過してもデータの**平均値や分散**
(ばらつき具合)が一定である状態のこと。



時系列分析では**定常性を満たしていないとモデル化できない**

自己共分散をもう少し説明します

自己共分散は、
同一時系列データの異時点間の共分散

$$\text{Cov}(y_t, y_{t-1}) = E[(y_t - \mu)(y_{t-k} - \mu)]$$

M_t t時点までの条件付平均
M_{t-k} t-k時点までの条件付平均

自己相関

$$\begin{aligned} \text{自己相関} &= \frac{\text{(自己共分散)}}{(\text{t時点の分散}) \times (\text{t-k時点の分散})} \\ &= \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t) * \text{Var}(y_{t-k})}} \end{aligned}$$

自己相関を日本語で説明します

自己相関とは、当日以前のデータの影響をどの程度受けているかを表す

- ・1日前と大きな自己相関がある場合

昨日の値が大きかった → 今日も多いだろうな

- ・3日前と大きな負の自己相関がある場合

3日前の値が大きかった → 今日は少ないだろうな

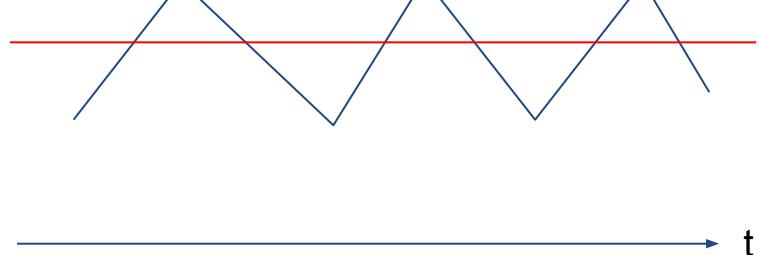
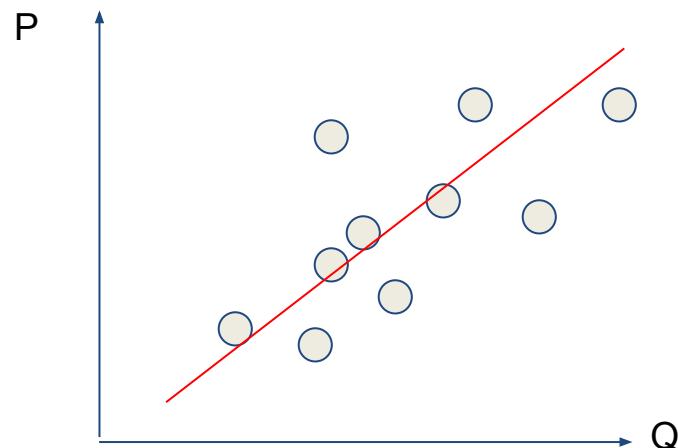
これまでの相関との違い

これまで

自己相関

異なる変数間の相関

同一変数の異時点の相関



例：t時点のデータはt-1時点のデータの影響がある

自己相関の特徴

$$-1 \leqq (\text{相関係数}) \leqq 1$$

corr

2変数間の相関

corr > 0 → 正の相関

Aが増えればBも増える

corr < 0 → 負の相関

Aが増えればBが減る

自己相関

corr > 0 → 正の相関

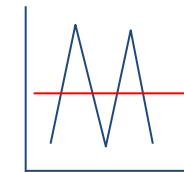
t時点の値 $y_t > M$ なら

時間が経つと共に大きくなる

corr < 0 → 負の相関

1時点前(t-1)とは逆の動きをしやすい

(※1次の場合)



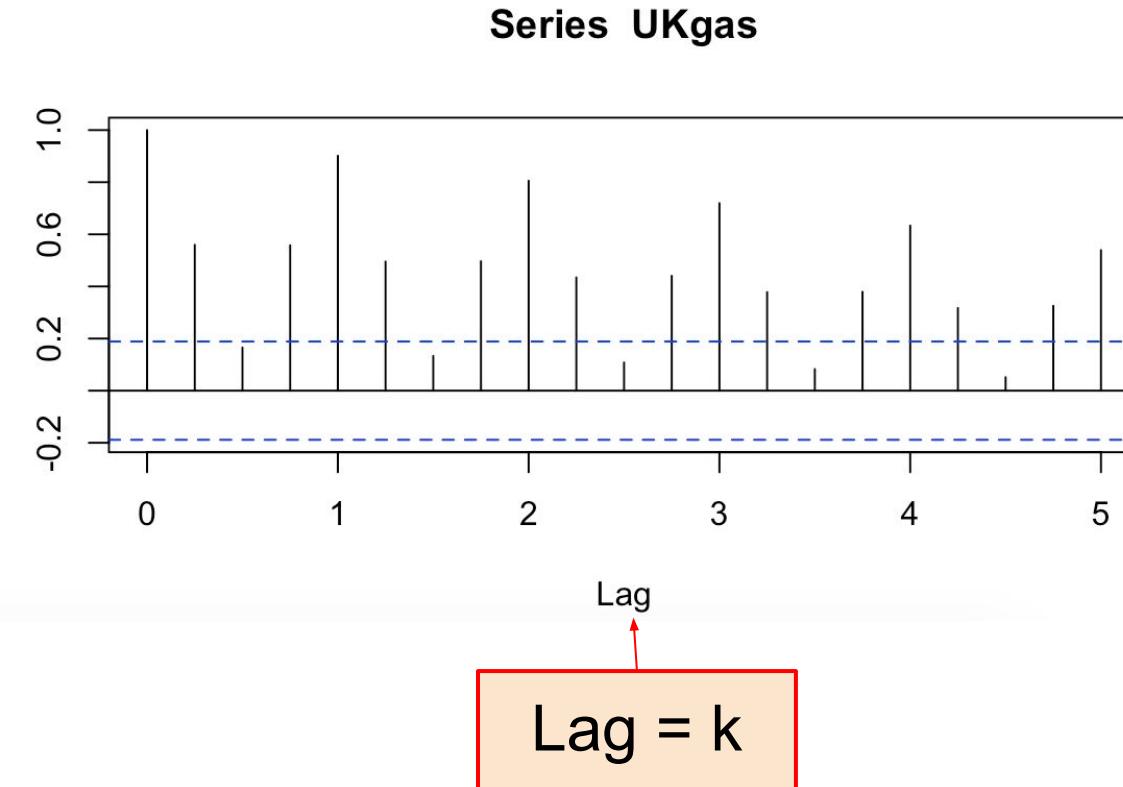
時系列データの特徴を捉えるための指標

- 標本自己相関係数
 $ACF = Auto\ Correlation\ Function$
- 標本偏自己相関係数
 $PACF = Partial\ Auto\ Correlation\ Function$
 - 「Partial = 偏」で偏微分の「偏」と同じ意味
 - PACFは、 y_t, y_{t-k} の単純な相関ではなく、その間の $y_{t-1}, y_{t-2}, \dots, y_{t-k+1}$ が y_t, y_{t-k} へ与える影響を取り除いたもの（推定方法は数式が大変なので割愛）

自己相関の可視化: コレログラム

ラグK($K=1,2,3\dots$)の自己相関をグラフに描いたもの

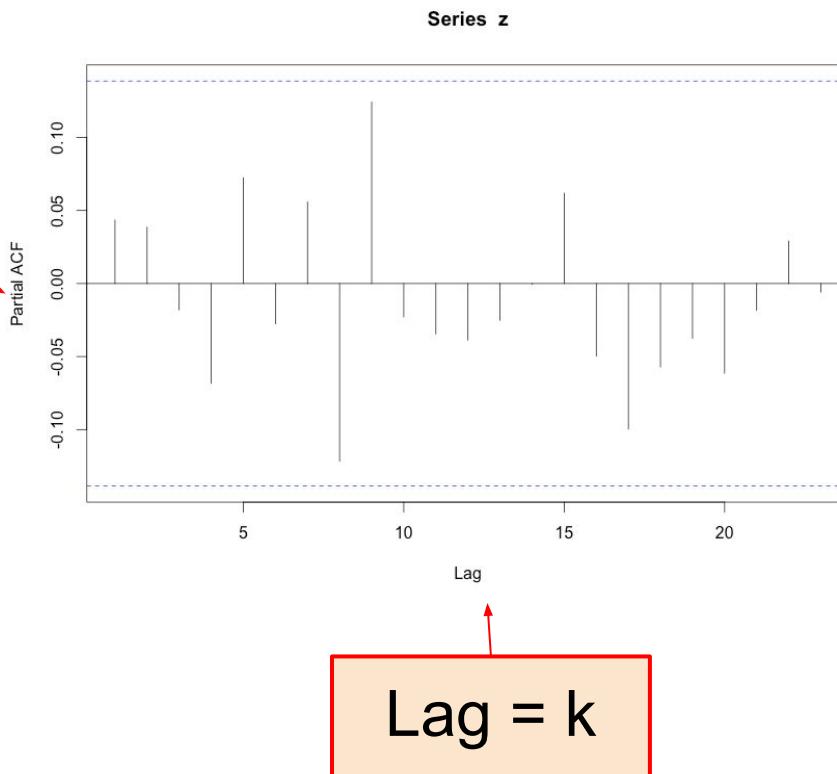
```
> acf ( data )
```



偏自己相関係数

ラグK(K=1,2,3...)の偏自己相関をグラフに描いたもの

```
> pacf ( data )
```



Rでやってみましょう

さきほど作成したWhite noiseの時系列データの自己相関係数、偏自己相関係数をプロットしてみましょう。

どのようなことに気が付きますか？

自己相関の検定

自己相関があるかどうかを調べるためにLjung-Box検定で確認を行う

Ljung-Box検定

帰無仮説: 自己相関がない

対立仮説: 自己相関がある

```
> Box.test( data , type="L" )
```

Ljung-Box検定の実行例

```
> Box.test(ukgas, type="L")
```

Box-Ljung test

data: ukgas

X-squared = 34.772, df = 1, p-value = 3.707e-09



帰無仮説が棄却されるので、自己相関があると判断できる

Rでやってみよう

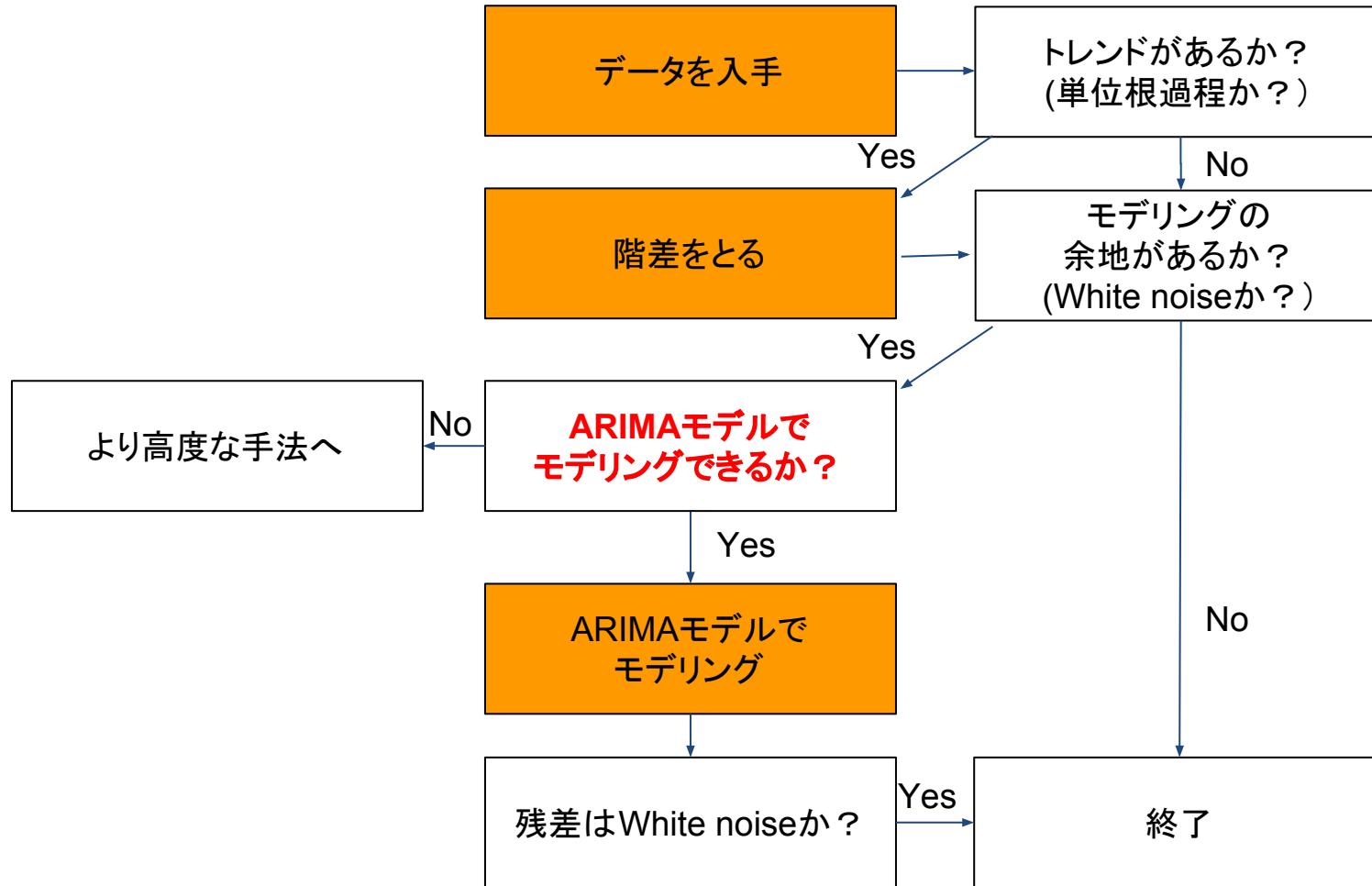
さきほど作成したWhite noiseをBox.testしてみましょ
う。



第4章

ARモデルとMAモデル

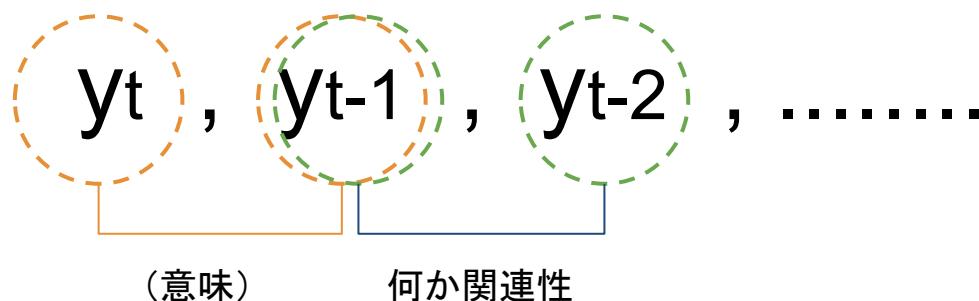
時系列分析の手順



時系列モデル

順序に見られる意味

→ 相関を見つけ出し表現するもの



直前、あるいは直前のk個の観測値が何かしら影響がある。
あるいは、同一の構成要素を持つ。

時系列分析の基本モデル

①AR(Auto Regressive)モデル

→自己回帰モデル

②MA(Moving Average)モデル

→移動平均モデル

③ARMA(Auto Regressive Moving Average)モデル

→自己相関移動平均モデル

AR(自己回帰)モデル

直前、あるいはk個前までの観測値が何かしらの影響を与えるようなモデル

※回帰モデル: $y = ax + b$ (y は x の影響を受ける)

自己回帰モデル: $\underline{y_t} = \underline{ay_{t-1}} + \underline{\varepsilon_t}$

直前のみの1次の自己回帰

y_t は直前の y_{t-1} の影響を受ける

※ 2次なら $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$

※ k次なら $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_k y_{t-k} + \varepsilon_t$

移動平均モデル(MA)

(過程)

直前、あるいはk個前までの観測値と
同様の構成要素をもつ。

(イメージ)

$$\begin{aligned}y_t &= a + b \\y_{t-1} &= b + c \\y_{t-2} &= c + d\end{aligned}$$



1次(直前のみ)

$$\begin{aligned}MA(1) = y_t &= M + \varepsilon_t + \theta_1 \varepsilon_{t-1} \\y_{t-1} &= M + \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \\y_{t-2} &= M + \varepsilon_{t-2} + \theta_3 \varepsilon_{t-2}\end{aligned}$$

※一般的な「移動平均」とは異なる

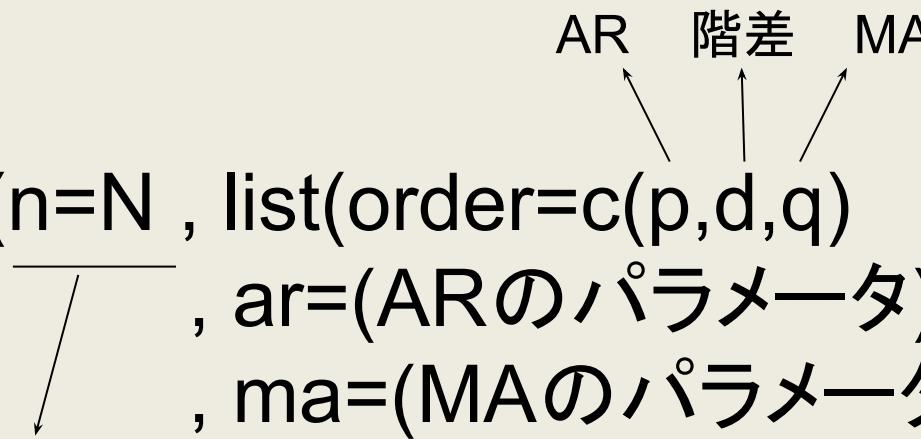
ε_k :確率変動するもの

Rでやってみよう

- arima.sim関数を使ってAR過程、MA過程のデータを作成し、acf関数、pacf関数を使って傾向を確認しましょう。
- 講師がホワイトボードに例題を書きますので、arima.simでデータを生成してみてください。

arima.sim(n=N , list(order=c(p,d,q))
 , ar=(ARのパラメータ)
 , ma=(MAのパラメータ))

観測数
(sample size)



演習

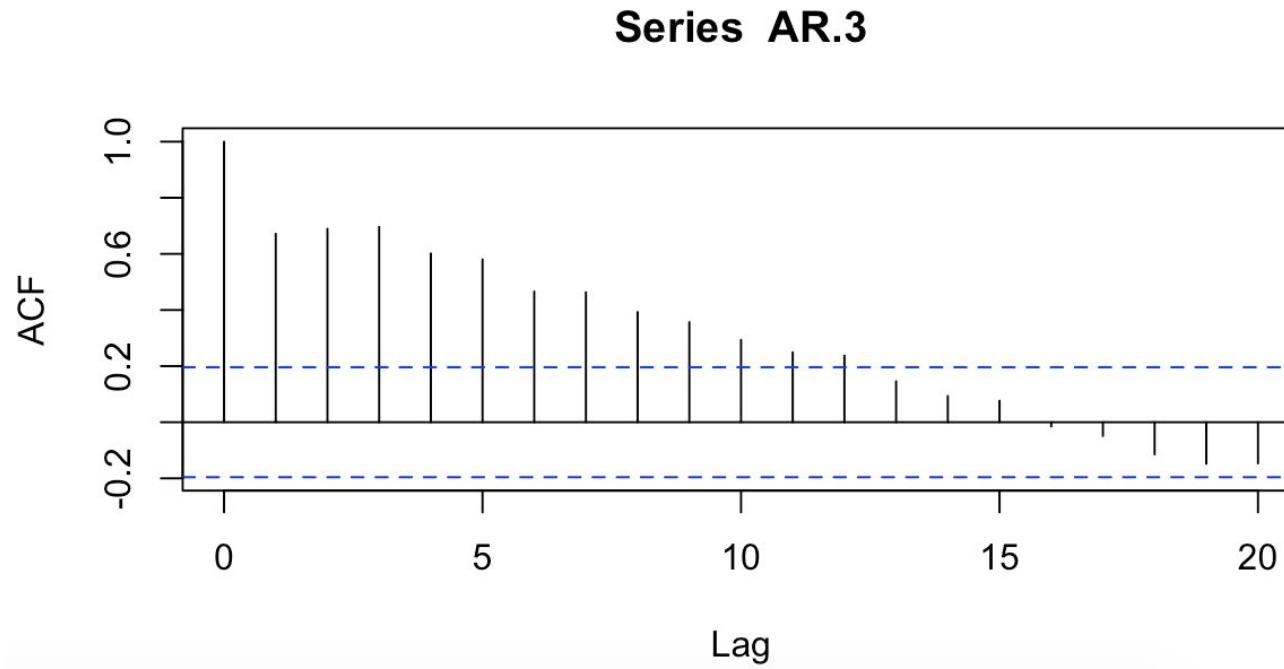
生成したデータについて、ACF、PACFをプロットしてみてください。どのような特徴がありますか？

ヒント

- Lagが増えていくと、徐々に遞減するでしょうか？
- それとも突然ゼロに近くなるでしょうか？

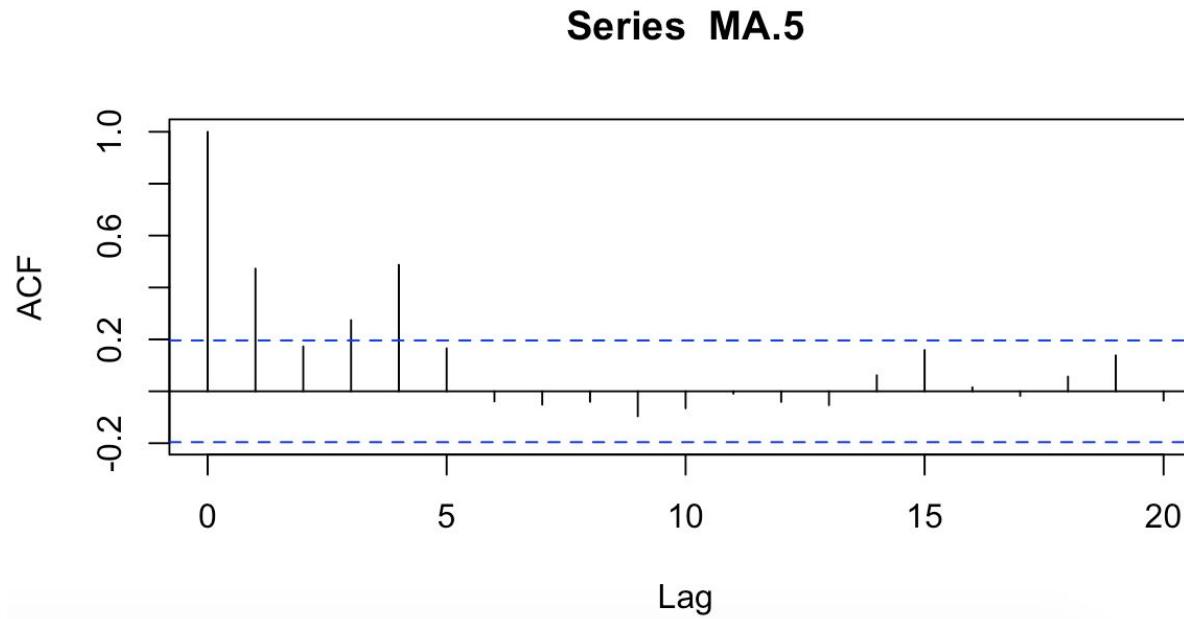
ARモデルの特徴

自己相関係数の絶対値は徐々に減少していく



MAモデルの特徴

- ・MA(q)過程の $q+1$ 次以降の自己相関は0になる



ARとMAの関係

以下のように1次のARモデルがあります。

$$y_t = a_1 y_{t-1} + \varepsilon_t \dots \textcircled{1}$$

当然、 $t-1$ のモデルは以下のようになりますが、

$$y_{t-1} = a_1 y_{t-2} + \varepsilon_{t-1} \dots \textcircled{2}$$

①の y_{t-1} に②を代入してみてください。

同様に、 $y_{t-2}, y_{t-3}, y_{t-4} \dots$ と続けていくとどうなるでしょうか？

ARはある条件のときを満たすときだけ定常

- MAモデルは常に定常
 - whiteノイズは定常
 - whiteノイズの線形和であるMAモデルもやはり定常
- ARモデルがMAモデルに変形できれば、そのARモデルもやはり定常だと言える
 - ただし、係数の絶対値が1よりも小さいときのみ
 - そうでないときは、非定常
 - つまり、ARモデルは定常とは限らない

上記は1次のARモデルの話ですが、一般化すると
「特性方程式」(この講義では触れません)と呼ばれる方程式の解の絶対値が1よりも小さいときにARモデルは定常となる(第5章と関係します)。

話を戻して… ARとMAモデルを合体させる

AR(自己回帰)モデル

+

MA(移動平均)モデル

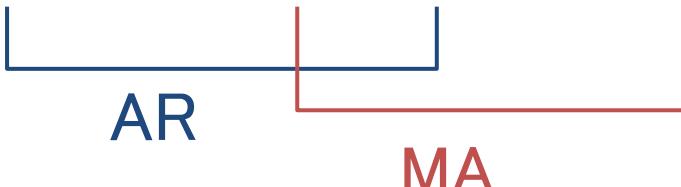


自己回帰移動平均モデル

ARMAモデル

ARとMAの両性質を併せ持つモデル

1次のARMAモデル

$$\text{ARMA}(1) \quad y_t = a_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$


Rでやってみよう

ARMAモデルのデータを生成してみましょう。

その上で、

- ACF関数
- PACF関数

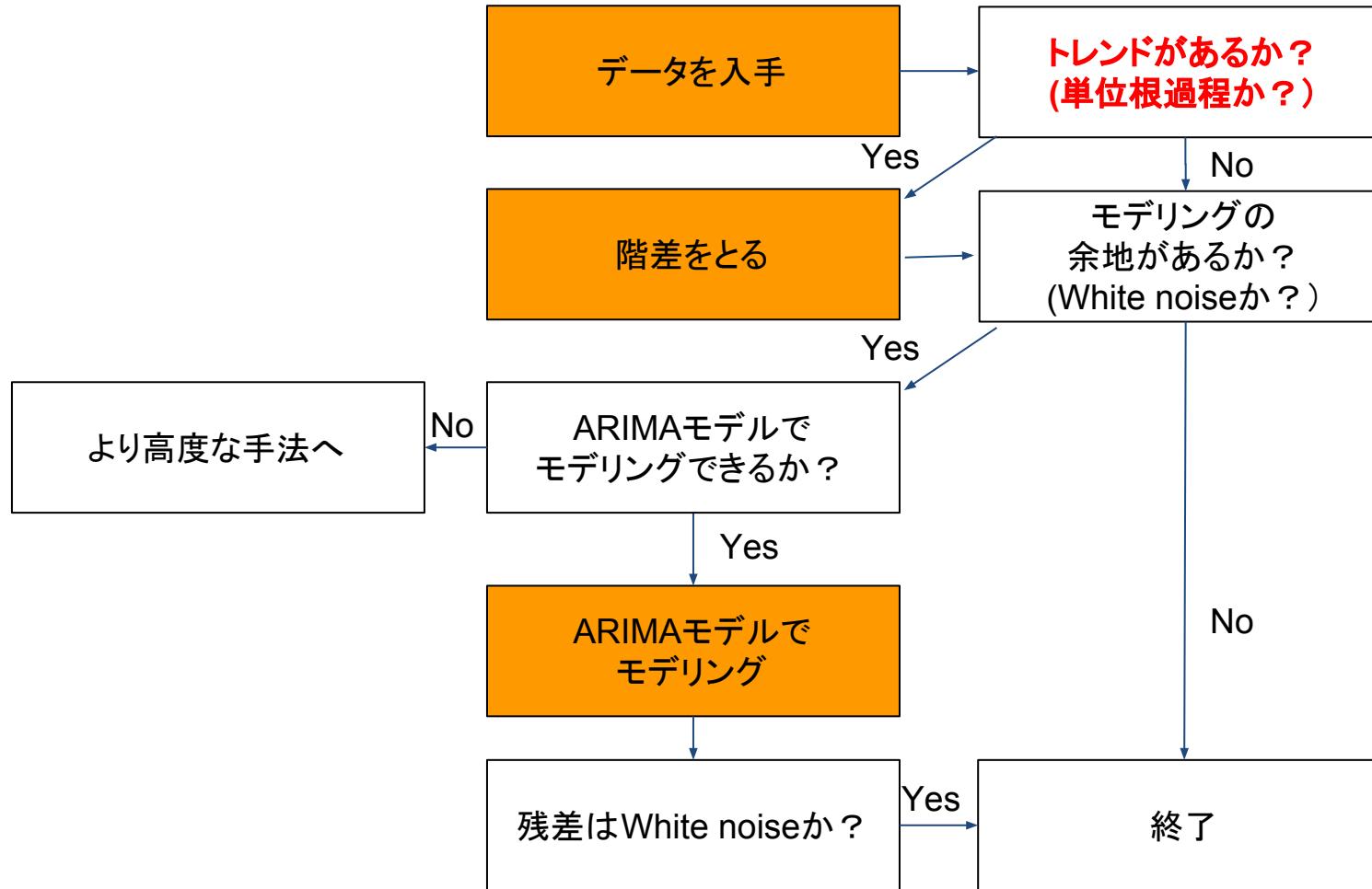
を描いて、特徴を確認してみましょう。



第5章

トレンドと単位根過程

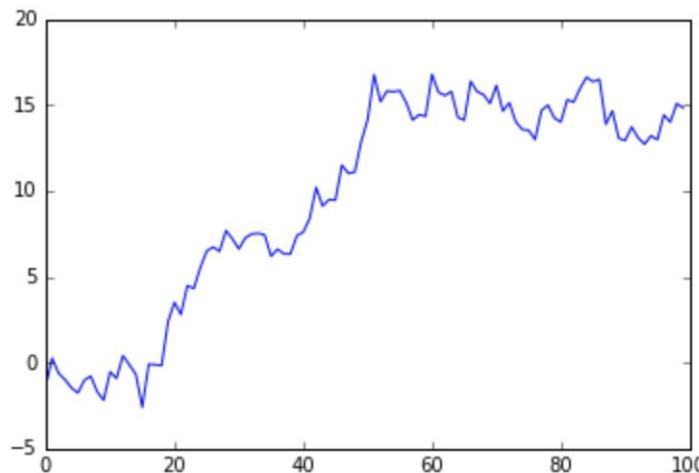
時系列分析の手順



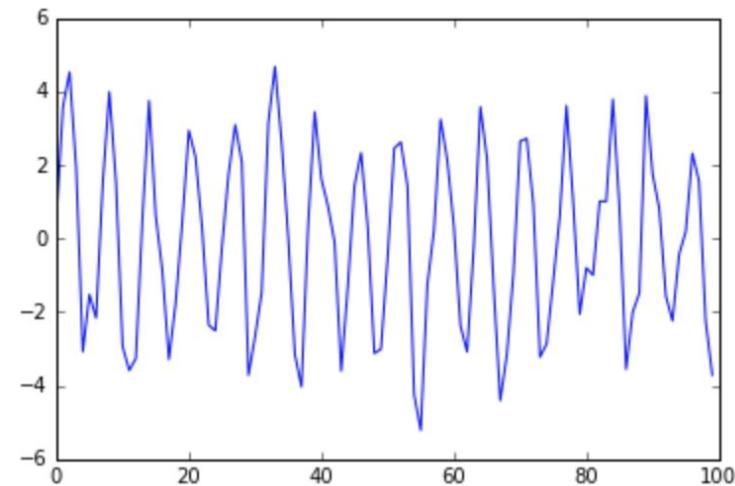
トレンドと季節性

実際に分析するデータは非定常過程の場合も多い。非定常過程になってしまう大きな要因として以下の2つが考えられる。

トレンド



季節性(周期性)



ランダムウォーク: Rでやってみましょう2

ランダムウォークで時系列データを作ってみましょう。

<ランダムウォークの式>

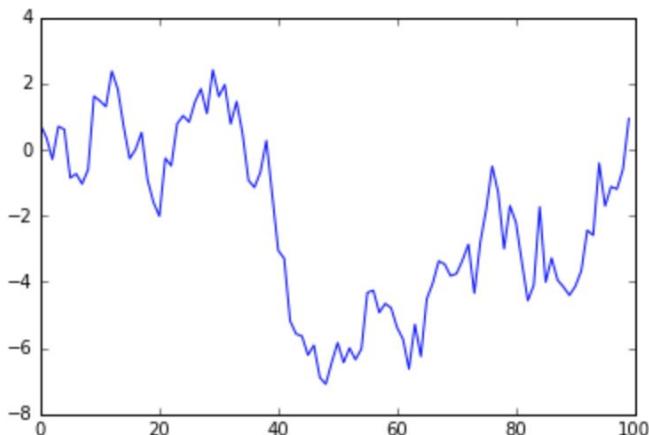
$$Y_{\{t\}} = Y_{\{t-1\}} + \varepsilon \quad \varepsilon \sim N(0, \sigma)$$

ヒント

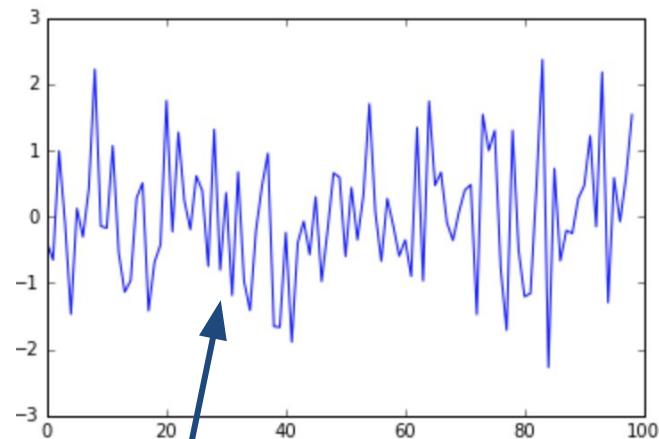
- $Y_{\{0\}} = 0$ で良いです。

単位根過程

元の過程が非定常で、その過程の差分系列が定常過程であるとき単位根過程という。単位根過程の代表例にランダムウォークがある



単位根過程



定常過程

メジャーな単位根検定は2種類ある

ランダムウォークを見抜くには
「単位根検定(Unit root test)」で
判定を行う。

Augmented Dicky Fuller検定(ADF.test)
または
Phillips-Perron検定(PP.test)

単位根検定

過程が定常過程かどうかを調べる検定

帰無仮説: 単位根AR(p)過程である

対立仮説: 定常AR(p)過程である

帰無仮説が棄却されればその過程は定常過程

```
> PP.test ( data )
```

Rでやってみよう

さきほど作成したランダムウォークの系列について、
単位根検定を実施してみよう。

単位根過程の実行例

```
> PP.test(x)

Phillips-Perron Unit Root Test

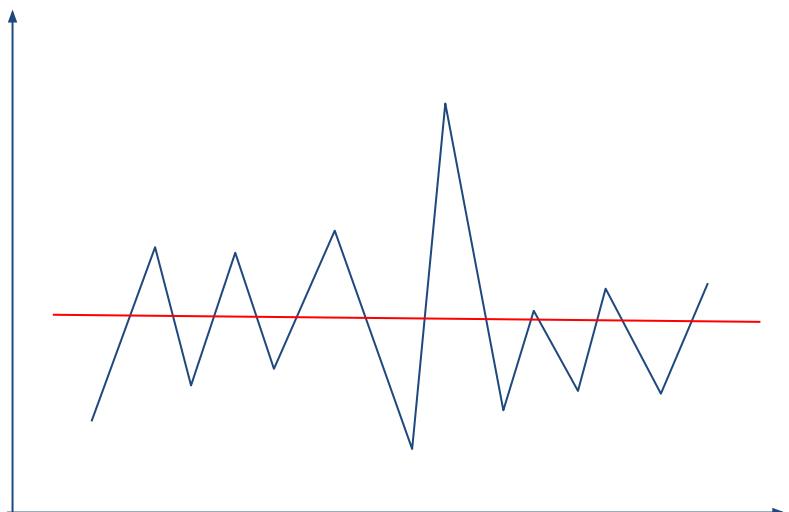
data: x
Dickey-Fuller = -12.591, Truncation lag parameter = 4,
p-value = 0.01
```

p-valueに注目すると、帰無仮説は棄却される
すなわち定常過程

単位根検定で棄却できた意味は？

検定した時系列データが

単位根過程でない



『定常過程』である

時系列分析で扱うための前提条件

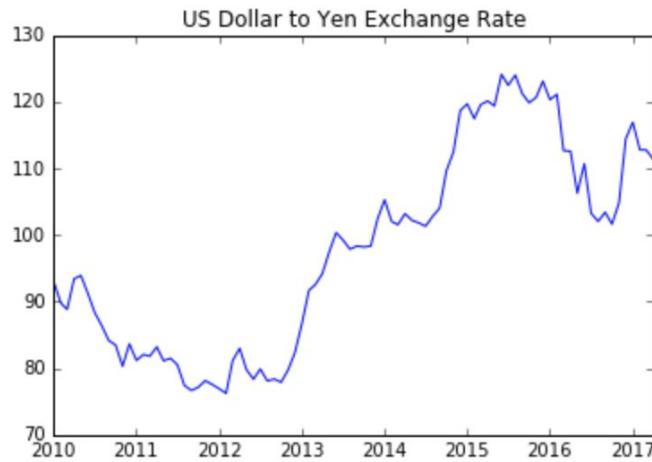
Rでやってみよう 2つにランダムウォークを回帰してみましょう

2つのランダムウォークのデータを作り、片方からもう一方を回帰してみてください。

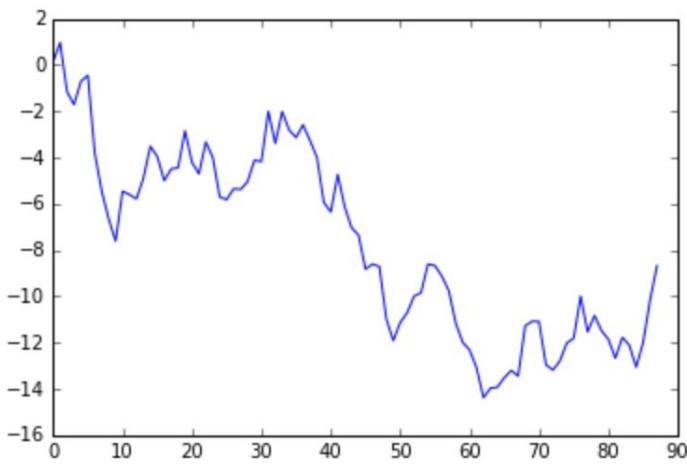
何が起こるでしょうか？

見せかけの回帰

単位根過程同士を回帰させると、本来存在しないはずの関係が生じてしまう



為替の時系列データ

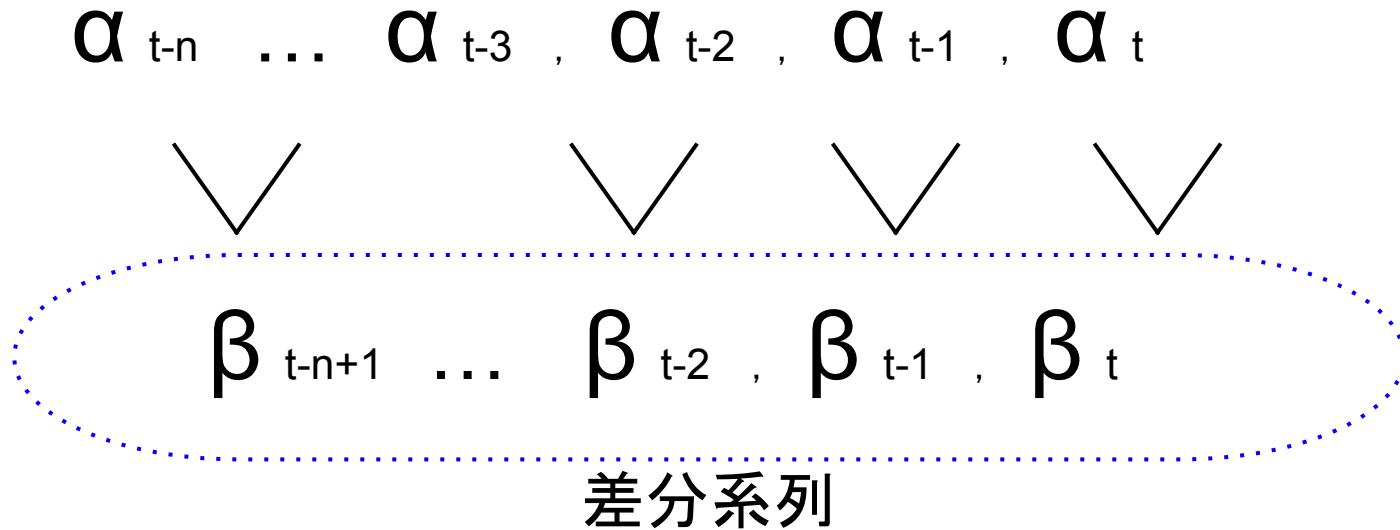


ランダムウォーク(単位根過程)

時系列データをランダムウォークで回帰させると…

話を戻して、
どうすればよいでしょうか？

時系列データの階差をとった
「差分系列」を作る



ARIMA モデル？

自己回帰 移動平均 モデル
Auto regressive Moving Average

和分 Integrated

和分？

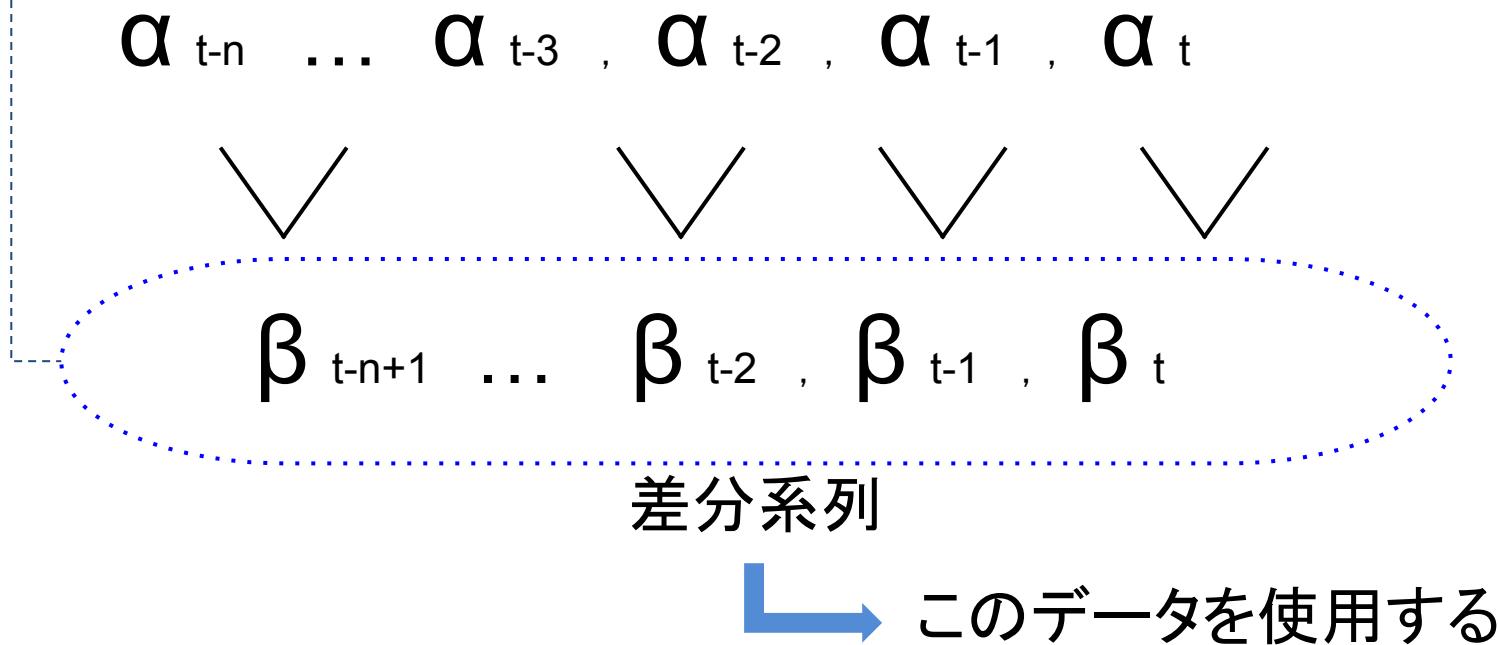


階差を使ったもの

階差を使った？

※おさらい

時系列データの階差



ARIMAの活用例

株価、為替レートといった単位根過程の
時系列データをモデルに落とし込む

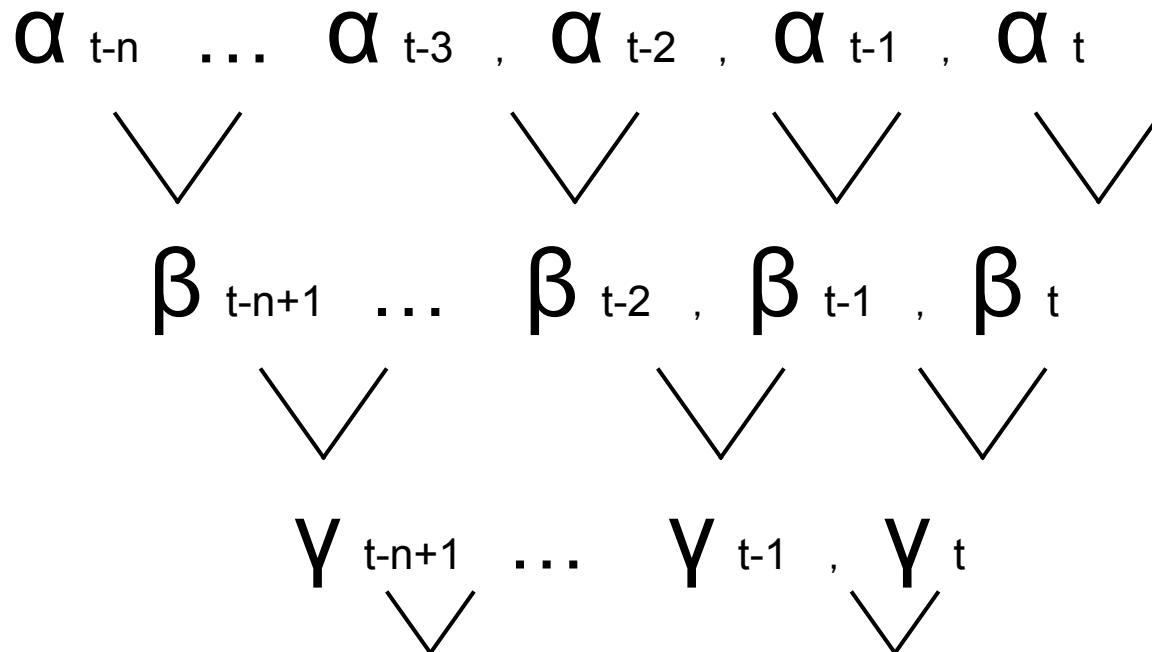


ARMAモデルに差分系列を付け加える
(Integrated)

階差は1回とは限らない

1回のみでは非定常の場合もある

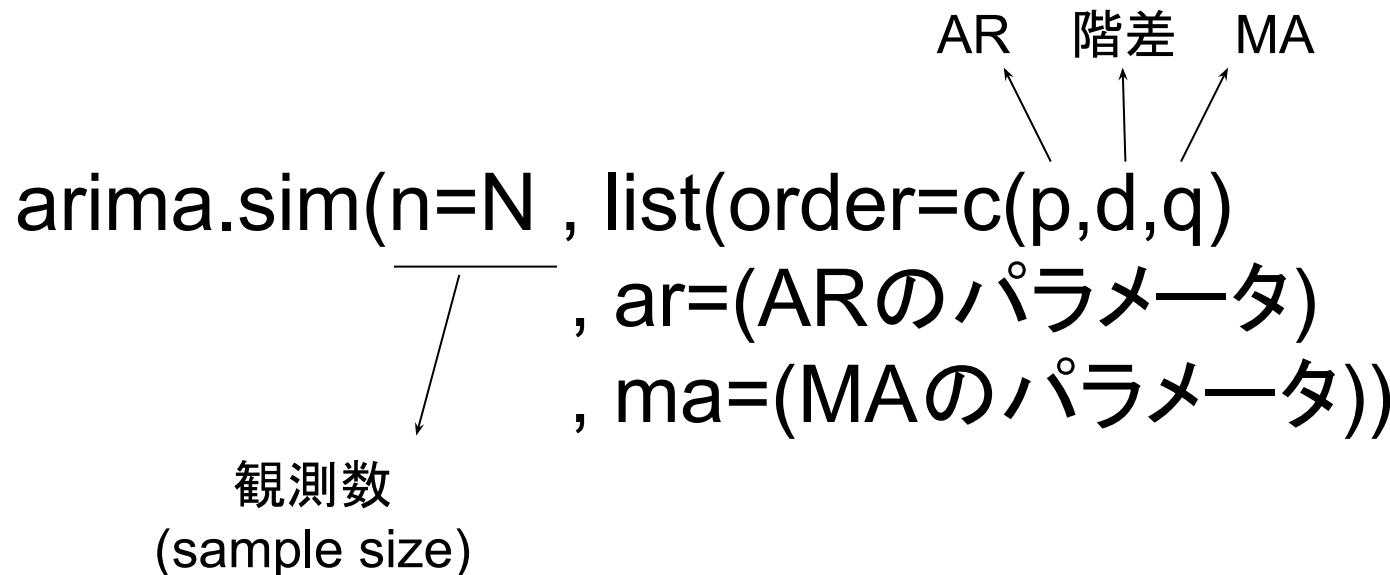
階差後の「差分系列」が定常過程になるまで行う



Rでやってみよう: ARIMA過程の生成

AR(1), MA(1), 階差1で、ARIMA過程のデータを生成してみましょう。その上で、

- 時系列プロット
- Box.test, ACF関数、PACF関数
- 単位根検定
- 階差を取ったうえで上記を実施



最適な次数はどうやって調べればよいか

パラメータ数の比較には...

定番の AIC

でも1個ずつ調べなくては…？

No! RではプログラミングしなくともAICを網羅的に比較する関数がある

```
> library(forecast)  
> auto.arima( データ ,  
             max.p = P , max.q = Q, max.d = D  
             , stepwise = T , trace = T )
```

max.p = ARモデルの次数上限

max.q = MAモデルの次数上限

d = 差分の階数→ 単位根過程では必須

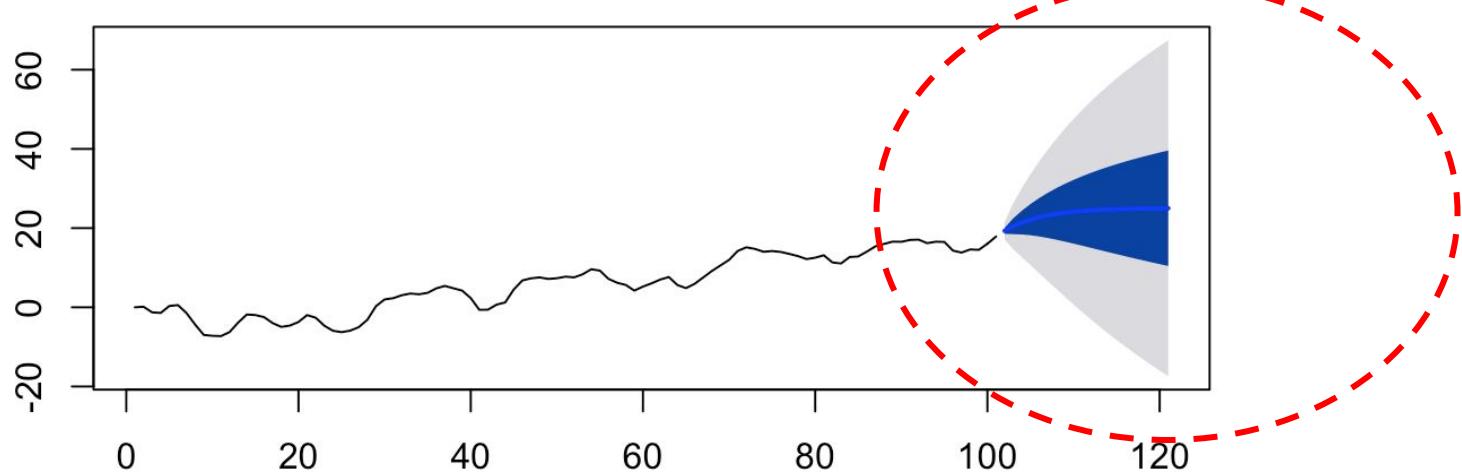


設定した条件下でAICが最も小さいものを選ぶ

ARIMAモデルで予測してみる

```
> library ( forecast )
> plot ( forecast ( model , level = c ( x , y ) , h = term )
```

level : 信頼区間を設定
h : 予測区間



Rでやってみよう: ARIMAの推定と予測

さきほど生成したデータが与えられたと過程して、

- モデルの推定
- 予測

をやってみましょう。



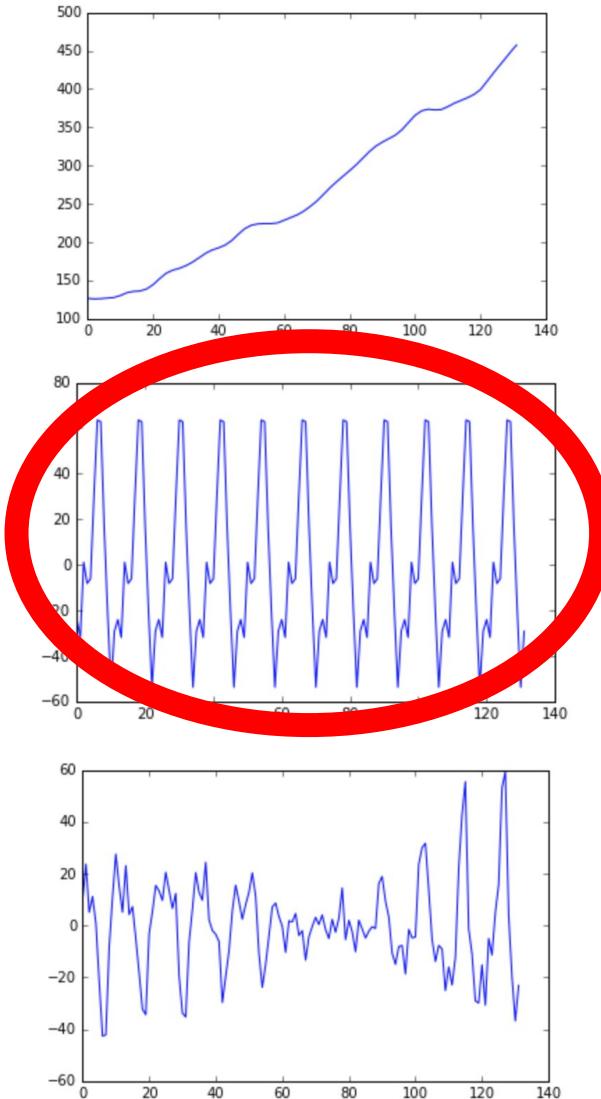
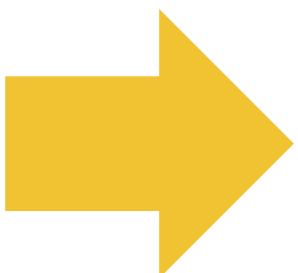
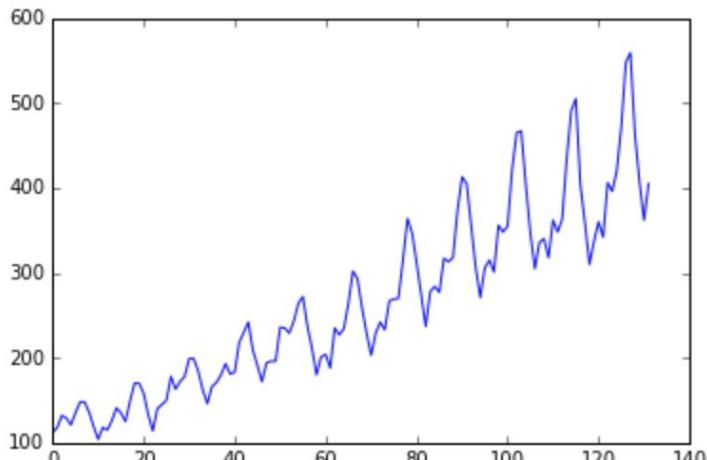
第6章

季節調整済時系列モデル

(SARIMAモデル)

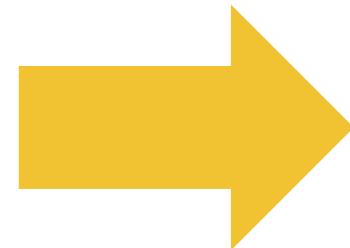
季節性とはどういうもの？

季節性 = 周期性



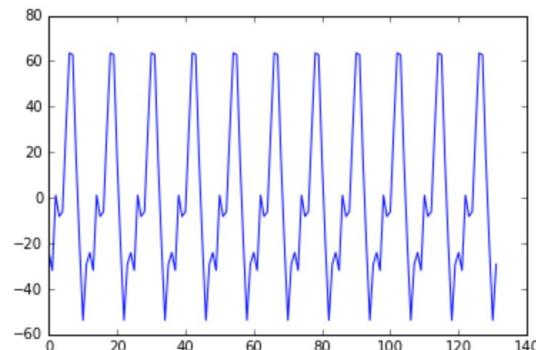
*おさらい 時系列分析の基礎は定常性

季節性などの
周期的な動き

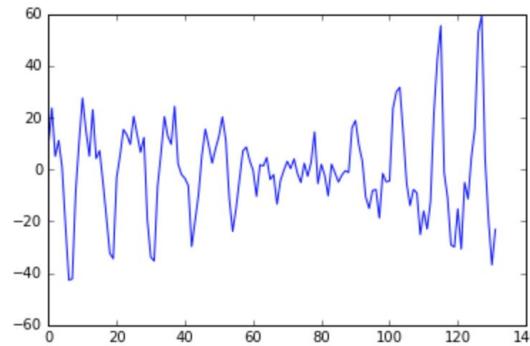


取り除いて
定常性を確保

周期



その他



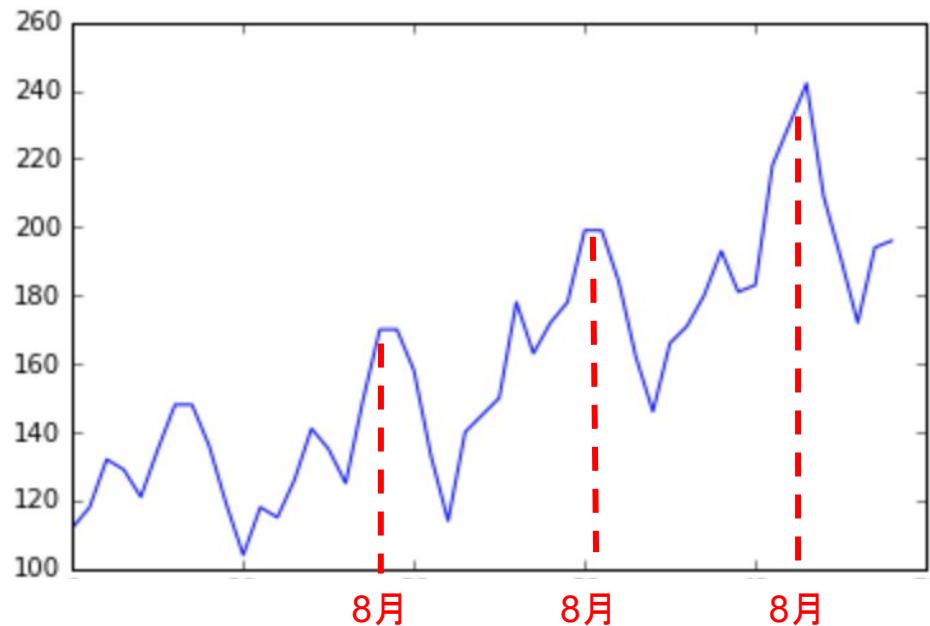
分離してモデル化



季節性・周期性を取り除く

(単位根過程同様) 階差をとろう

ex. 前年同期比(月別データなら12ヶ月前と階差)



(Y年1月) - (Y-1年1月)
■
■
(Y年8月) - (Y-1年8月)
■
■
(Y年12月) - (Y-1年12月)

季節性のあるデータの時系列モデル

S ARIMS モデル
seasonal



ARIMAモデルを使いつつ、
季節性も取り入れる

SARIMAモデルの考え方

ARIMA → $\left\{ \begin{array}{l} \text{トレンド除去} \\ \text{単位根過程} \end{array} \right\}$ → 階差を入れる

AR + I + MA

SARIMA → 季節階差を取り入れる

→ $\left[\begin{array}{l} \text{AR} \\ \text{I} \\ \text{MA} \end{array} \right]$ の $\left[\begin{array}{l} \text{どれか} \\ \text{or} \\ \text{全て} \end{array} \right]$ 季節階差を入れる

どこに組み入れたら良いか？

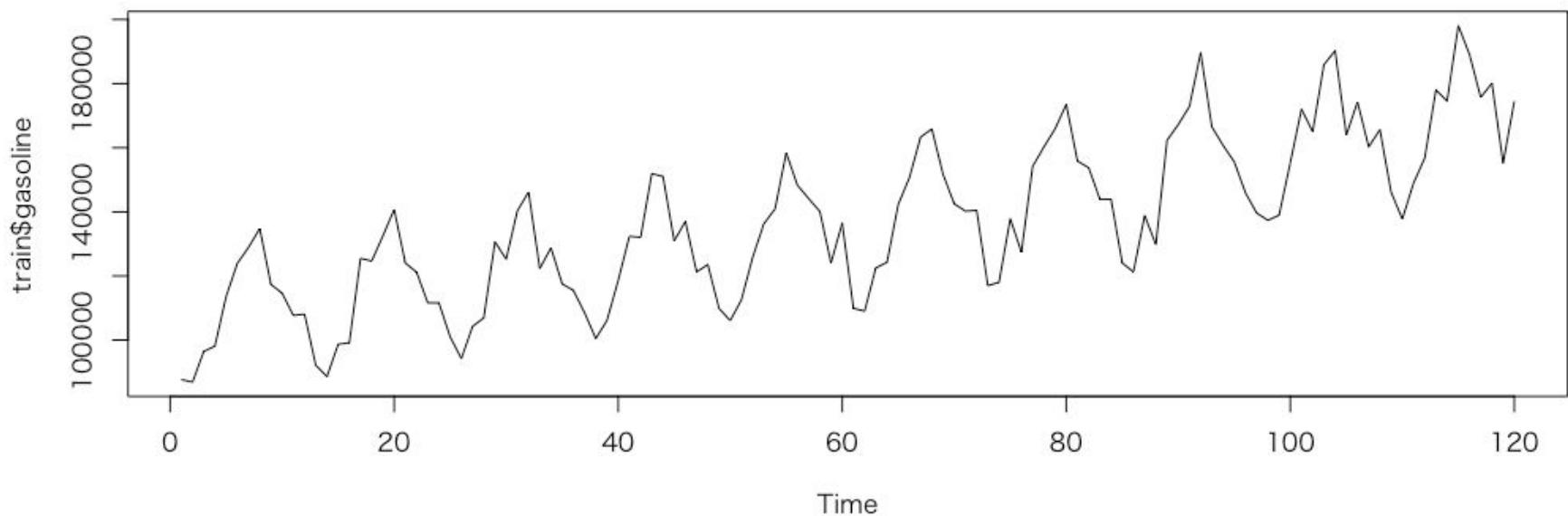
心配ご無用!!



(周期性さえ見つけられれば)
Rが勝手に判断してくれます

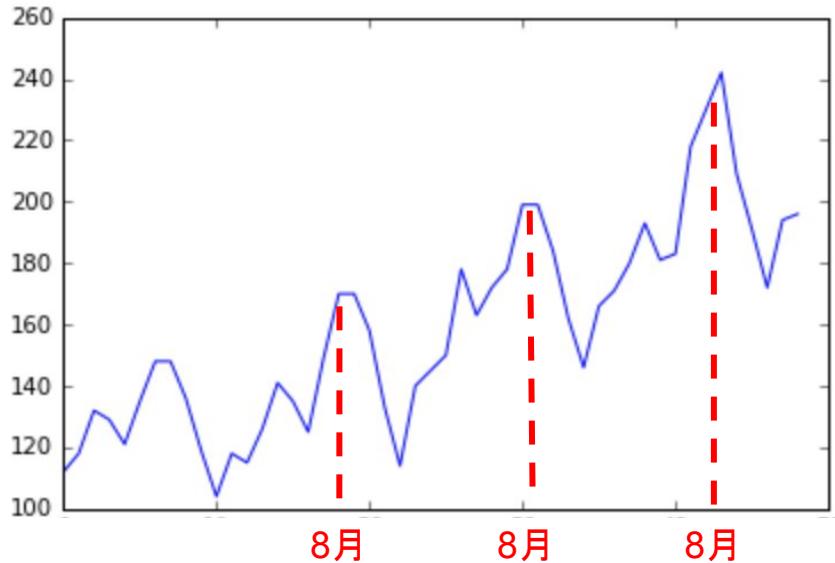
Rでやってみよう

gas_train.csvとgas_test.csvを読み込み、周期性を見つけ出してください。



Rで周期性を指定するには

データを見る

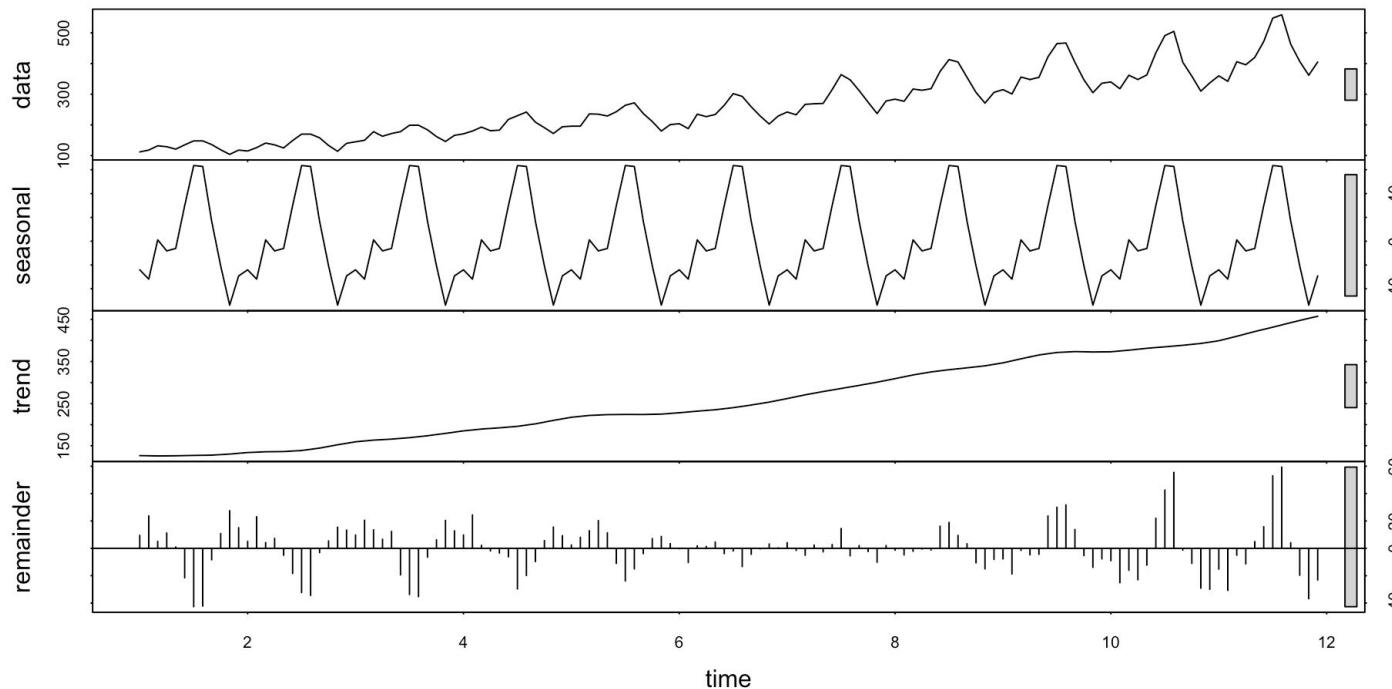


なるほど...
月別データで
1年の周期性がある

```
> ts ( data , frequency = 12 ) -> data.ts
```

時系列の要素分解

```
> library ( stats )
> stl ( data.ts , s.window = " periodic " ) -> ts.stl
> plot ( ts.stl )
```



Rでやってみよう

auto.arima関数でモデルを推定してみましょう。

ヒント

- トレンドがあるのか？（あるなら階差とる）
- ホワイトノイズではないよね？
- AR, MAの最大次数はどのくらい？

```
> library ( forecast )
> auto.arima (data.ts , max.p = P , max.q = Q , seasonal = T)
```

SARIMAモデルを推定

```
> library ( forecast )
> auto.arima (data.ts , max.p = P , max.q = Q , seasonal = T)
-> ts.sarima
```

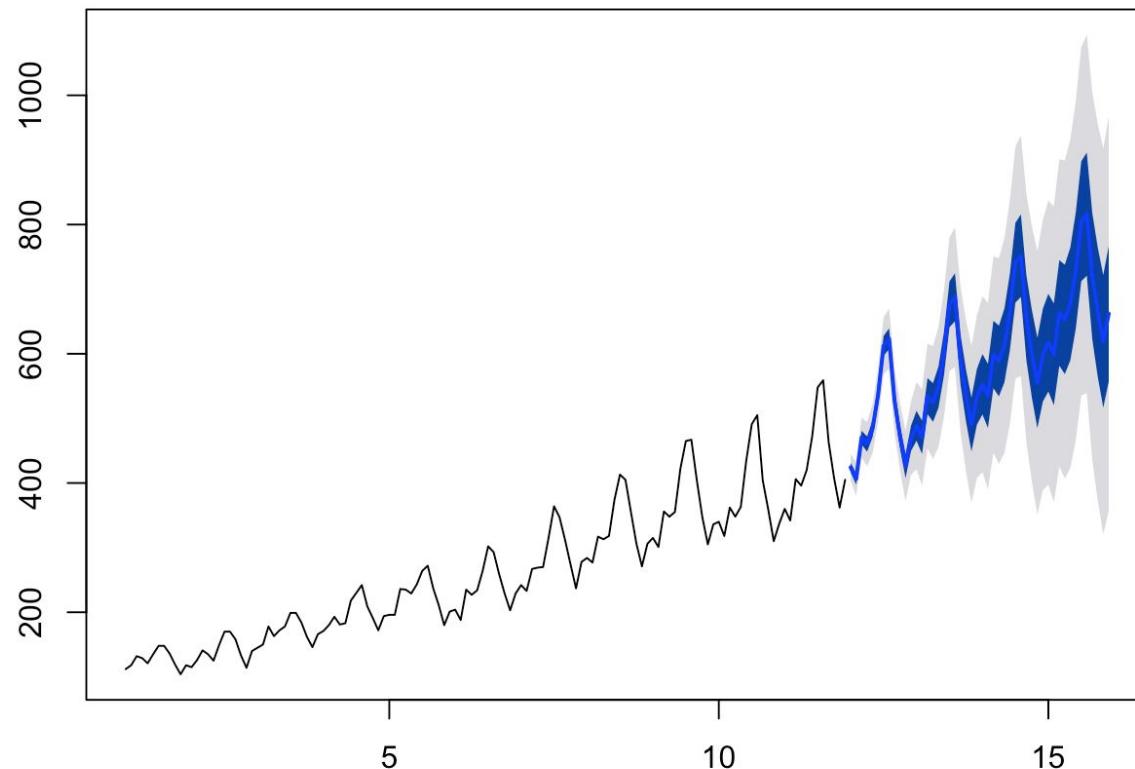
ARIMA(2,1,1)(2,0,0)[12]	: 579.1967
ARIMA(1,1,1)(2,0,0)[12]	: 590.3084
ARIMA(2,1,0)(2,0,0)[12]	: 589.2675
ARIMA(2,1,2)(2,0,0)[12]	: 581.3342
ARIMA(1,1,0)(2,0,0)[12]	: 593.1792
ARIMA(2,1,1)(2,0,0)[12] with drift	: 579.6145

Best model: ARIMA(2,1,1)(2,0,0)[12]

周期12でARに季節2階差が
最適なモデル

SARIMAモデルで予測

```
> plot(forecast ( ts.sarima , range = c ( x, y ) , h = term ))
```





実践 日本経済を予測してみよう

やってみよう!!

日本の失業率(1972年7月~2011年12月)の
データを使って以後3ヶ月(36ヶ月)の失業率
を予測してみよう

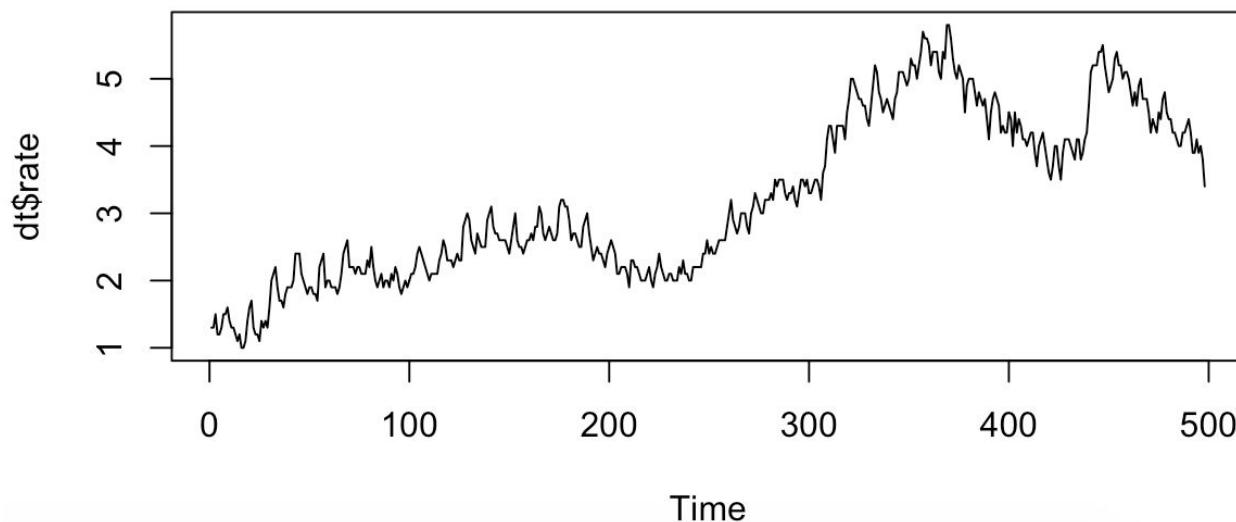
1. データを読み込んでプロット

```
#ファイル読み込み
```

```
> read.csv ( "unemploy.csv" , header = T ) -> dt
```

```
#プロットしてみる
```

```
> plot.ts ( dt$rate )
```



2.そのまま扱ってよいか確認

```
#単位根過程(ランダムウォーク)でないか検証  
#ライブラリーをロード  
> library ( tseries )  
> PP.test ( dt$rate )
```

```
> PP.test(dt$rate)
```

Phillips-Perron Unit Root Test

data: dt\$rate

Dickey-Fuller = -2.1821, Truncation lag parameter = 5, p-value = 0.5012

p-value = 0.5012

棄却できない



ランダムウォーク

3. 階差をとて定常過程に

#1階差を取る

```
> diff ( dt$rate , 1 ) -> dt.diff
```

#階差した結果で再度検定

```
> PP.test ( dt.diff )
```

```
> PP.test(dt.diff)
```

Phillips-Perron Unit Root Test

data: dt.diff

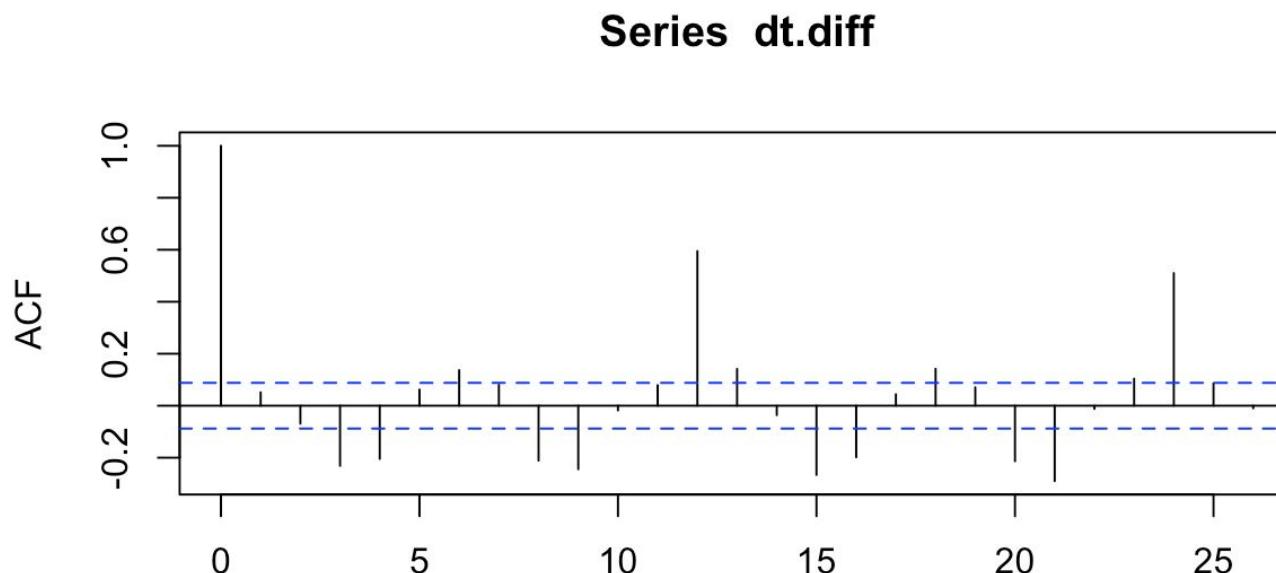
Dickey-Fuller = -21.342, Truncation lag parameter = 5, p-value = 0.01

1階差でOK!



4. 階差をとつて差分過程の自己相関の有無を確認

```
#自己相関を可視化  
##予め関数をライブラリで用意  
> library ( stats )  
##自己相関を見る  
> acf ( dt.diff )
```



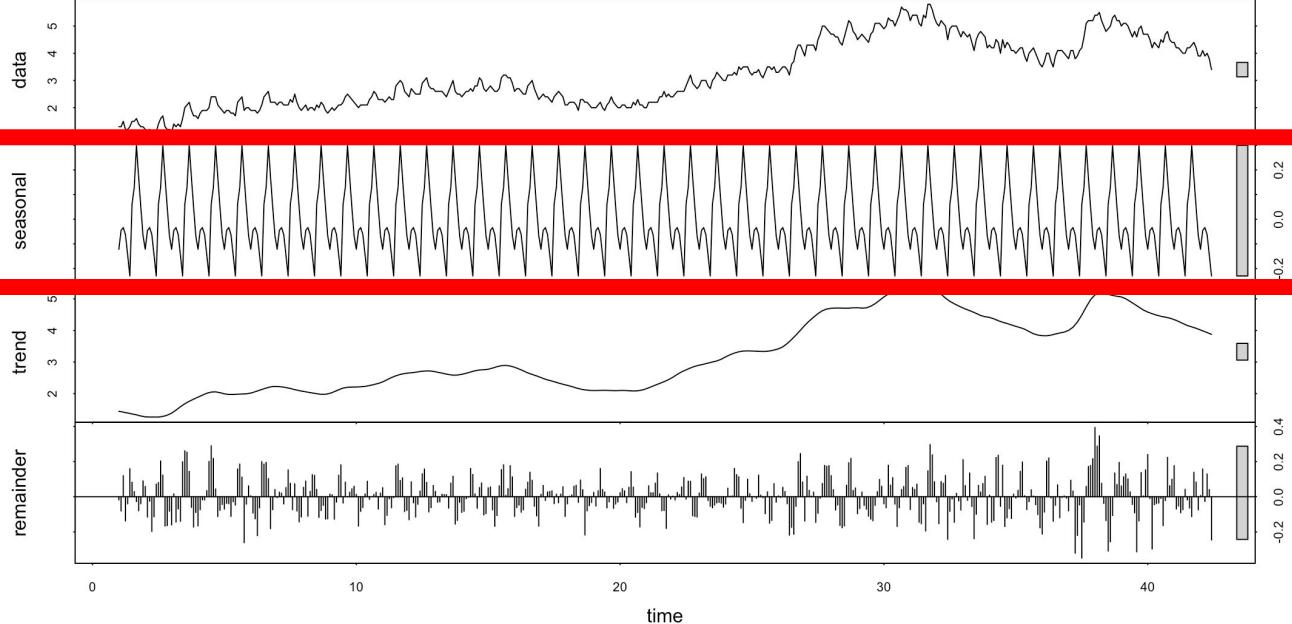
5.季節性(周期性)の確認

#月単位のデータとして読み込んでみる

```
> ts ( dt$rate , frequency = 12 ) -> dt.ts
```

#周期性の確認

```
> plot ( stl ( dt.ts , s.window = " periodic " ) )
```



季節性あり

6.モデル・次数の選択

#AR?MA?ARIMA?また何期前のデータを使うか？

#最適なモデル(AICが最も小さいモデル)を探す

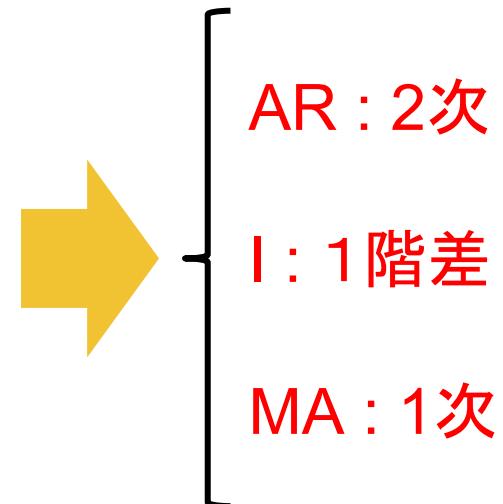
> library (forecast)

> auto.arima (dt.ts , trace = T , stepwise = T
, seasonal = T) -> dt.arima

ARIMA(3,1,2)(2,0,0)[12] with drift	:	-577.28
ARIMA(2,1,1)(2,0,0)[12]	:	-580.176
ARIMA(2,1,1)(1,0,0)[12]	:	-536.4599
ARIMA(2,1,1)(2,0,1)[12]	:	Inf
ARIMA(1,1,1)(2,0,0)[12]	:	-569.8473
ARIMA(3,1,1)(2,0,0)[12]	:	-579.8517
ARIMA(2,1,0)(2,0,0)[12]	:	-575.9341
ARIMA(2,1,2)(2,0,0)[12]	:	-579.8419
ARIMA(1,1,0)(2,0,0)[12]	:	-569.6373
ARIMA(3,1,2)(2,0,0)[12]	:	-579.3072

季節調整

Best model: ARIMA(2,1,1)(2,0,0)[12]



6.モデル・次数の選択

#AR?MA?ARIMA?また何期前のデータを使うか？

#最適なモデル(AICが最も小さいモデル)を探す

> library (forecast)

> auto.arima (dt.ts , trace = T , stepwise = T
, seasonal = T) -> dt.arima

ARIMA(3,1,1)(2,0,0)[12]	:	-579.8517
ARIMA(2,1,0)(2,0,0)[12]	:	-575.9341
ARIMA(2,1,2)(2,0,0)[12]	:	-579.8419
ARIMA(1,1,0)(2,0,0)[12]	:	-569.6373
ARIMA(3,1,2)(2,0,0)[12]	:	-579.3072

Best model: ARIMA(2,1,1)(2,0,0)[12]

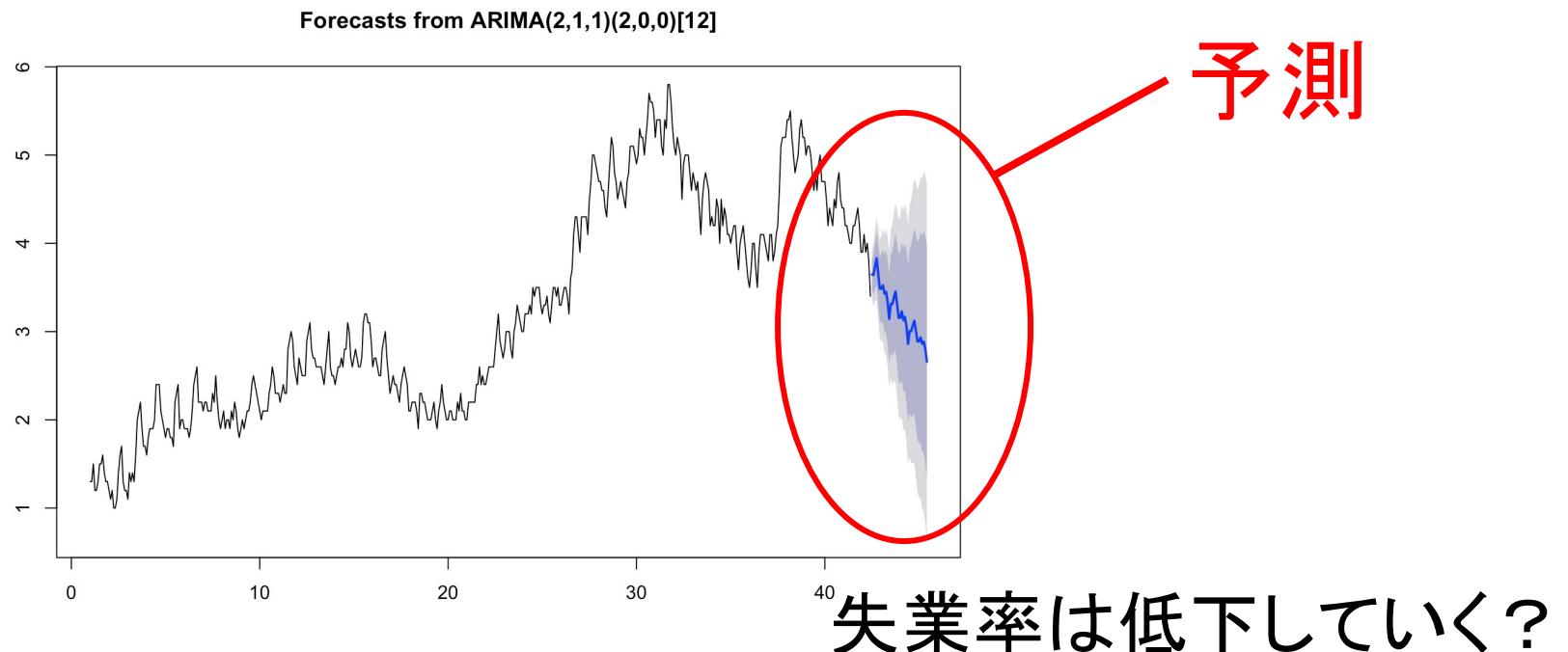
季節調整

AR : 2次 | : 1階差 MA : 1次

7. 予測

#プロットしてみる

```
> plot ( forecast ( dt.arima , level = c( 50 , 95 ) , h = 36 ) )
```





第7章

VARモデル

VARモデル

VARモデル: ベクトル自己回帰モデル

Vector Auto Regressive model

ある時系列データの推定モデルをつくるのに
自己を含む複数の時系列データを用いる方法

ex. ベクトル =

$$\begin{pmatrix} a_{11}, a_{12}, a_{13}, \dots, a_{1j} \\ a_{21}, a_{22}, a_{23}, \dots, a_{2j} \\ \vdots \\ a_{i1}, a_{i2}, a_{i3}, \dots, a_{ij} \end{pmatrix}$$

何のために？

①複数の時系列データを用いることで
单一のときよりも予測精度を上げたい

②複数の時系列データのうちどの変数が影響あるのかを知りたい

例えばどんな時に利用する？

- ・国際株価市場の関連性
(アメリカ株、欧洲株が日本株に与える影響)
- ・各経済指標が為替に与える影響
ドル円→失業率・消費者物価指数、GDP、etc

VARモデルをRで使ってみよう

次数を決定 (各情報量基準(AICなど)で最も小さいときの次数を出す)

```
> library(vars)  
> VARselect ( data , lag.max = p )
```

次数の上限

モデルを生成

```
> VAR ( data , p = P )
```

次数。※VARselectで出した最適な次数

VARで予測

生成したモデルでN期先を予測

```
> predict ( VARモデル , n.ahead = N , ci = P)
```

何期先まで予測するか

信頼区間

データ間の因果関係をとらえる

- ① グレンジャー因果
- ② インパルス応答関数
- ③ 分散分解

グレンジャー因果

単一の時系列データで予測した時よりも、もう1つの時系列データの双方でモデル化したときに予測精度が向上するかを測る

ex. 失業率を予測

- ①失業率単一の時系列データで将来予測
- ②失業率に加え、平均時給の両者で将来予測



比較して②のほうが予測精度が高い



平均時給は失業率に影響がある

グレンジャー因果の問題点

予測精度の向上→因果があることはわかるが、
影響度および原因 \leftrightarrow 結果の関係はみえない



インパルス応答関数で測る

インパルス応答関数

ある時系列データに変動があった場合に、**以後**
その他の時系列データにどの程度影響するか

ex. 失業率と平均時給の因果と影響度を予測

- ① 平均時給が変動(上昇／低下)した場合に、その後、失業率にどう影響するか
(あるいはその逆かどうか因果関係を測る)

分散分解

ある時系列データの将来の値を予測する場合に、
その他の時系列データがどの程度寄与するか

ex. 失業率と平均時給の将来的な関係性を予測

- ① 将来の失業率を予測する際に、予測できない変動の何%が平均時給に寄与するのか
(あるいは失業率自体が何%寄与するのか)を測る
- ② 将来の平均時給に対する失業の寄与度の場合も同様に測ることができる

VARモデルの演習

R studioを起動し、実践してみましょう。

時系列分析で学んだモデル

- ・ 時系列データ(定常性・扱いの注意)
- ・ 時系列モデル1(AR, MA, ARMA)
- ・ 時系列モデル2(ARIMAモデル, SARIMAモデル)
- ・ 時系列モデル3(VARモデル)



第8章

機械学習を使った時系列データ予測

ランダムフォレストを使う場合

ランダムフォレスト:

kaggleでも時系列データ予測に頻繁に使用されている

時系列データのもつ以下の特徴

- 「前の値が、次の値に影響を与える」
- 「周期性・季節性」

を、「特徴量の設計」で間接的に表現する。

特徴量の設計例

- 年月日 →

- 年・月・日をそれぞれ分解(連続値として保持)

- 週フラグ(1～50週に分けて、ダミー変数を取る)

- 曜日フラグ(日～月に分けて、ダミー変数を取る)

- 時間 →

- 1時間ごとに分解(0～23に分けダミー変数を取る)

- 3時間ごとに分解(例:0～2, 3～5, …, 21～23)



総合演習

課題

Walmartの店舗売上予測に取り組んでください。

- 最終スコア
- 使用したモデル
- 作成した特徴量

The screenshot shows the competition page for "Walmart Recruiting - Store Sales Forecasting" on Kaggle. The page features the Walmart logo and a brief description: "Use historical markdown data to predict store sales" with "691 teams - 4 years ago". The navigation bar includes tabs for Overview, Data, Discussion, Leaderboard, Rules, Team, My Submissions, and Late Submission (which is highlighted). The Overview section contains a "Description" box stating: "One challenge of modeling retail data is the need to make decisions based on limited history. If Christmas comes but once a year, so does the chance to see how strategic decisions impacted the bottom line." It also lists "Evaluation", "Prizes", and "Timeline". Below this is a large text block about the competition rules, mentioning historical sales data for 45 Walmart stores, departments per store, and selected holiday markdown events like Clearance, Rollbacks, and Special Buys. It also notes that participants must compete individually and can contact hosts for interviews.

<https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting>