

L13 時系列データ

樋口さぶろお <https://hig3.net>

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習 L13(2022-01-06 Thu)

最終更新: Time-stamp: "2022-01-06 Thu 08:00 JST hig"

今日の目標

- 多変量データと時系列データの違いを説明できる
- 階差, 累積, 移動平均, 自己相関係数, コレログラムの意味を説明できる. Pandas で求められる.



L12-Q1 L12-Q2

Quiz 解答:ブートストラップ法による標準誤差

- ① $\frac{1}{20-1}[(8-11)^2 + \cdots + (12-11)^2] = \frac{24}{19}$. 標準誤差の推定値は $(24/19)^{1/2}$.
- ② 0.05, 0.95 分位点は 8, 12. よって, $10 - (12 - 10) < \theta < 10 + (8 - 10)$.

上は、上端下端までの長さが同じという特殊な状況になっちゃったので、もうちょっと一般的な状況の間と答.

L12-Q3

TA Prob and Sol: ブートストラップ法による標準誤差と信頼区間

(理解チェックのための不自然なブートストラップ標本です)

標本で、 θ の推定量 $\hat{\theta} = 9$ となった. $B = 20$ ブートストラップ標本で、推定量 $\hat{\theta}^*$ は、大きさの順に、

8, 8, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12 となった.

- 推定量 $\hat{\theta}$ の標準誤差を推定しよう.
- θ の $\alpha = 0.1$ の基本的ブートストラップ信頼区間を求めよう.

略解

- $\frac{1}{20-1} [(8-11)^2 + \dots + (12-11)^2] = \frac{24}{19}$. 標準誤差の推定値は $(24/19)^{1/2}$.
- 0.05, 0.95 分位点は 8, 12. よって, $9 - (12 - 9) < \theta < 9 - (8 - 9)$.

ここまで来たよ

12 ブートストラップ法

13 時系列データ

- 時系列データ
- 移動平均
- 自己相関係数

時系列解析 Time Series Analysis

時系列 時間 t に依存する量の列 $x(0), x(1), x(2), \dots, x(t), \dots$
 以前の値が、今の値に影響. $x(t)$ $t = 0, 1, 2, 3, \dots$ は独立でない.

例

- 特定の銘柄の毎日の株価のデータ
- 週ごとの売上のデータ
- 1分おきの気温のデータ
- 1年ごとの太陽黒点の個数のデータ
- 時刻 t のランダムウォーカーの座標 $X(t)$



時系列解析 時系列を解析する手法群 経済統計学でさかん.

目的 時系列を再現する. $t \leq T$ のデータから $t > T$ を予測する.

標本 (データ) を解析 → **時系列モデルを作成** → 再現・予測

ランダムウォーク

確率モデル (3Q4)

時系列データの形式

多変量時系列データ $\mathbf{x}(t)$.

日時	気温	気圧
2021-12-18 Sat 10:00	10.1	1010
2021-12-18 Sat 11:00	10.5	1005
⋮	⋮	⋮

日時には実質的な意味がある。日時を消して順序を変えてはいけない。

日時 i のデータは日時 $j (< i)$ のデータの影響を受けることがある。

多変量データ

選手番号	100m 走	棒高跳び
1	10.1	2.30
2	10.4	2.12
⋮	⋮	⋮

選手番号は便宜上のもので、選手番号を振り直したデータも同じ意味。

(同じ分布からとってきたデータ点かもしれないが) 選手 i のデータが選手 j のデータの影響を受けるようなことはない。

時系列データ

Python

Dataframe df

1列だけ取りだしたものの `df['x']` (Series) で時系列を表せる

Series には日時の index t をつけられる (従来は整数のデータ通し番号だった)

数列 $a_t \leftrightarrow x(t)$

日時

日時は深い…

Pandas, Python では

日時を Index (選手番号のような, 変量でない通し番号) として扱う.

yyyy-mm-dd hh:mm:ss のような文字列を, DatetimeIndex 型に変換すると, 期間の合計, 期間の平均, 特定の月や曜日のみ, などの処理ができる.

`df['x'].resample('M').mean()` 週ごとの平均の数値を求める

'M'(月)の他に, 'W', 'D', 'Y' など. `resample()` のマニュアル参照.

平均 `mean()` の他に和 `sum()` など.

時系列データの計算

階差数列

`df['x'].shift(k)` = $x(t - k)$ ずらす演算

`df['x'].diff(k)` = $x(t) - x(t - k)$ ずらして差をとる演算

累積 cumulative sum

`df['x'].cumsum()` = $\sum_{t'=1}^t x(t')$.

増加率

$y(t) = x(t)/x(t - 1)$, $y(t) = x(t)/x(t - 1) - 1$.

時系列の3要素

以下では、等間隔、日時順に並んでいるデータだとして、日時 index を省いて考える

現実の時系列は次の3つの重ね合わせになっていることが多い

$$x(t) = \boxed{\text{トレンド}} + \boxed{\text{周期的変動}} + \boxed{\text{ランダム成分}}$$

- **トレンド** (長期的傾向) 期間を通して時間に比例して増減する傾向. 一過的な増減
- 短い周期の**周期的変動** 季節, 週, 月, 年

フーリエ級数解析, フィルタ (パターン情報処理)

- **ランダム成分** (ノイズ) 時刻ごとに独立な確率変数

定常な時系列

(標本ナントカ)の意味で、日時のどの一部分を取り出しても「本質的に」同じであるとき、**定常 (stationary)** という
トレンドがあれば非定常 (氷河期と今は同じでない)
周期的変動があるなら非定常 (夏と秋は同じでない, 昼と夜は同じでない)
トレンドや周期的変動を引いておくと定常になる. トレンドや周期的変動を移動平均で消すと定常になる.
近いうちに、母ナントカの立場で**定常**を明確に定義します.

ここまで来たよ

12 ブートストラップ法

13 時系列データ

- 時系列データ
- 移動平均
- 自己相関係数

移動平均 Moving Average

時系列 $x(t)$ から平滑化 (smoothing) した別の時系列 $y(t)$ を作る手法

$(2\ell + 1)$ 次の移動平均

$$y_{2\ell+1}(t) = \frac{1}{2\ell + 1} \sum_{t'=t-\ell}^{t+\ell} x(t')$$

$x(0)$ $x(1)$ $x(2)$ $x(3)$ $x(4)$ $x(5)$ $x(6)$ $x(7)$ $x(8)$ $x(9)$ $x(10)$ $x(11)$

2ℓ 次の移動平均

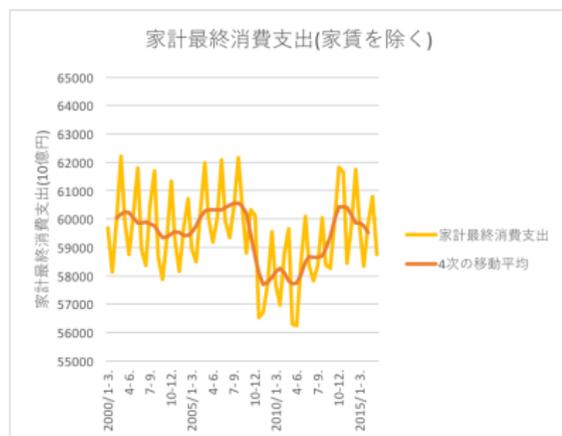
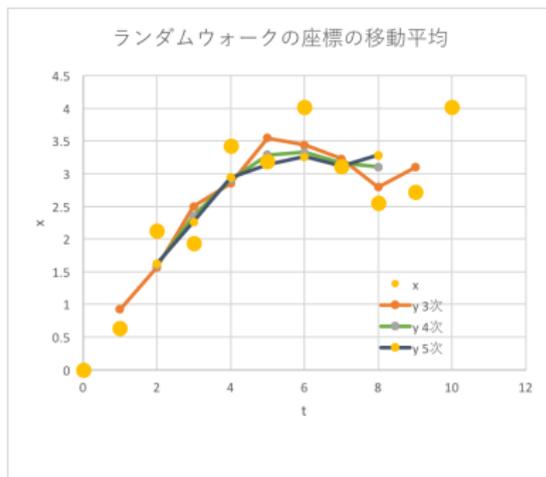
$$y_{2\ell}(t) = \frac{1}{2\ell} \left(\frac{1}{2} \cdot x(t-\ell+1) + x(t-\ell+2) + \cdots + x(t) + \cdots + x(t+\ell-2) + \frac{1}{2} x(t+\ell-1) \right)$$

$x(0)$ $x(1)$ $x(2)$ $x(3)$ $x(4)$ $x(5)$ $x(6)$ $x(7)$ $x(8)$ $x(9)$ $x(10)$ $x(11)$

離散フーリエ級数の第0項 [mva-d13-0-tsa.ipynb](#)

移動平均の性質

- 次数が高くなるほど滑らかになる
- 元のデータの「真ん中へん」を通る.
- 時間帯の端までは描けない



移動平均の性質 2

- **トレンド** (長期的傾向) 期間を通して時間に比例して増減する傾向. 一過的な増減 → 移動平均ではっきり見えるようになる
- 短い周期の**周期的変動** 季節, 週, 月, 年 → (周期くらいの) 移動平均で消える. (もとのデータ)-(移動平均) ではっきり見える フーリエ級数解析, フィルタ (パターン情報処理)
- **ランダム成分** (ノイズ) 時刻ごとに独立な乱数 → 移動平均で小さくなる

ラグ ℓ の後方移動平均 $y(t) = \frac{1}{\ell} \sum_{t'=t-\ell+1}^t x(t')$

ラグ ℓ の前方移動平均 $y(t) = \frac{1}{\ell} \sum_{t'=t}^{t+\ell-1} x(t')$

Python

```
df['x'].rolling( $\ell$ ).mean()
```

L13-Q1

Quiz(移動平均)

次の時系列データから、3,4 次の移動平均を求めよう.

t	0	1	2	3	4	5	6	7	8	9
x	1.8	-1.6	2.6	-1.2	3.0	-1.2	3.4	-0.8	3.8	0.0
y_3										
y_4										

ここまで来たよ

12 ブートストラップ法

13 時系列データ

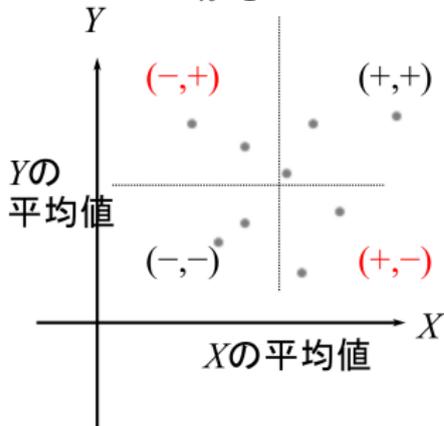
- 時系列データ
- 移動平均
- 自己相関係数

復習:標本共分散と標本相関係数

標本共分散 (covariance)

$$x, y \text{ の共分散 } C_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \times (y_i - \bar{y})$$

(不偏なら $\frac{1}{n-1}$)



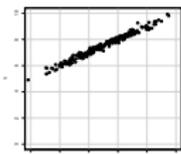
$(+, -) = (x_i - \bar{x} \text{ の符号}, y_i - \bar{y} \text{ の符号})$.

相関係数

$$\text{標本相関係数 } r = \frac{C_{xy}}{s_x s_y}$$

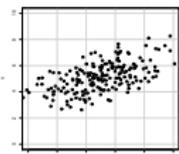
s_x, s_y : 標本標準偏差

分子分母で $\frac{1}{n}$, 不偏 $\frac{1}{n-1}$ をそろえる.



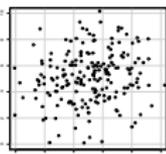
強い正の相関

$$r = 0.99$$



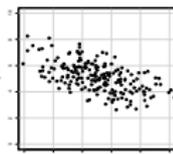
弱い正の相関

$$r = 0.55$$



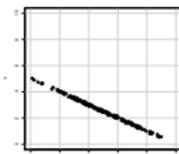
無相関

$$r = 0$$



弱い負の相関

$$r = -0.55$$



強い負の相関

$$r = -0.99$$

相関

‘正の相関’: x が大きい $\Leftrightarrow y$ が大きい

‘負の相関’: x が大きい $\Leftrightarrow y$ が小さい

強い/弱い: 傾向がはっきりしている/していない

標本自己共分散, 標本自己相関係数

k 次の標本自己共分散, 標本自己相関係数

時間 t を, ラグ k だけずらした $x(t), x(t - k)$ を 2 変量データだと思って, 標本共分散, 標本相関係数を考えたもの

$k = 1$ の例

x	y
$x(1)$	—
$x(2)$	$x(2 - 1)$
$x(3)$	$x(3 - 1)$
\vdots	\vdots
$x(T - 1)$	$x(T - 1 - 1)$
$x(T)$	$x(T - 1)$
—	$x(T)$

k の例

x	y
$x(1)$	—
\vdots	
$x(k + 1)$	$x(1)$
\vdots	\vdots
$x(T)$	$x(T - k)$
\vdots	
—	$x(T)$

k 次の標本自己共分散 autocovariance

$$C(k) = \frac{1}{T-k} \sum_{t=k+1}^T (x(t) - \bar{x})(x(t-k) - \bar{x})$$

$$\text{ただし標本平均値 } \bar{x} = \frac{1}{T} \sum_{t=1}^T x(t)$$

 k 次の標本自己相関係数 (標本自己相関関数) autocorrelation

$$r(k) = \frac{\text{自己共分散}}{\sqrt{\text{分散}}\sqrt{\text{分散}}} = \frac{C(k)}{C(0)}$$

$C(0)$ は $x(t)$ をサイズ T の標本と思ったときの分散.

コレログラム correlogram

横軸 ラグ k , 縦軸 k 次の自己相関係数 のグラフ.

- 多くの定常モデル (自己回帰モデルなど) では, k が大きいほど (遠い時刻ほど), 標本自己相関係数 $r(k)$ の絶対値は小さくなる.

Python tsa=Time Series Analysis 時系列解析

自己相関係数 `statmodels.tsa.stattools.acf(df['x'])`

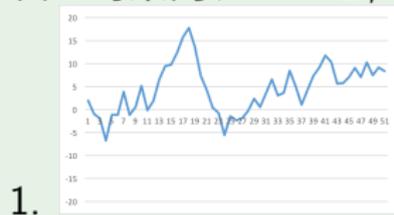
コレログラム `statmodels.graphics.tsa.plot_acf(df['x'])`

[mva-d13-0-tsa.ipynb](#)

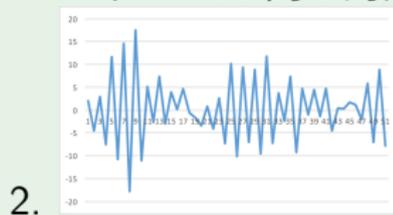
L13-Q2

Quiz(コレログラム)

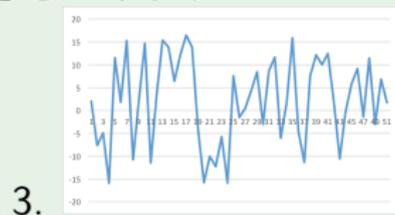
次の時系列データと、コレログラムの間の対応をつけよう。



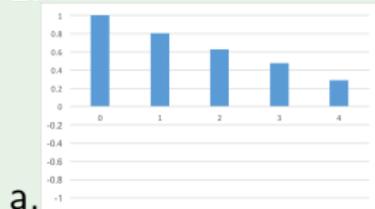
1.



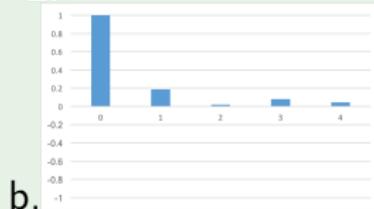
2.



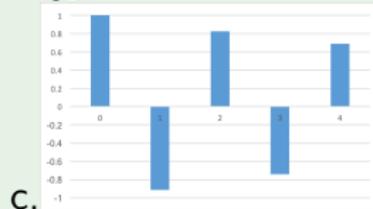
3.



a.



b.



c.

定常な時系列データに対する標本ナントカ

時系列分析では、定常過程について、1個の時系列データを、一定の長さに分割して複数個のデータからなる標本のように扱うことが行われる。

横: t (Pandas と逆), 縦:標本内のデータ番号.

標本自己相関係数を求めるとき、本当はこういう標本が欲しい。

	A	B	C	D	E	F	G	H	I	J
1	n \ t	0	1	2	3	4	5	6	7	8
2	1	2	-0.2522	-1.4293	0.41339	0.44925	0.30205	0.05638	1.16461	0.77765
3	2	2	0.40156	-0.1803	1.02945	1.31571	-0.3059	1.1405	0.1365	0.27561
4	3	2	1.74901	0.20238	-0.7695	0.67816	0.17627	0.66943	0.10917	1.30642
5	4	2	0.45349	-0.3762	-0.6185	0.45716	0.00561	-1.4453	0.09464	1.55283
6	5	2	1.01492	0.94089	-0.9599	0.44116	0.00072	0.5173	0.11447	1.38334
7	6	2	1.78265	1.70528	2.07552	0.06083	0.429	0.10902	-0.7164	-1.1645
8	7	2	0.69062	1.27761	1.15377	1.44955	-0.4663	0.24885	-1.323	-1.1454
9	8	2	0.74954	0.27013	1.05844	-0.1876	0.6692	-0.0232	1.09015	1.91198
10	9	2	1.21618	-0.2882	0.81639	1.84743	0.18484	-1.1622	-1.9981	-2.3211
11	10	2	1.20421	-0.487	-0.2274	0.34252	0.62865	1.46621	0.13736	0.13426

定常ならこれでもいいじゃん。Pandas の時系列を収める Series は $n = 1$ の横 1 行を縦に転置したもの。

	A	B	C	D	E	F	G	H	I	J
1	n \ t	0	1	2	3	4	5	6	7	8
2	1	2	-0.2522	-1.4293	0.41339	0.44925	0.30205	0.05638	1.16461	0.77765
3	2	2	0.40156	-0.1803	1.02945	1.31571	-0.3059	1.1405	0.1365	0.27561
4	3	2	1.74901	0.20238	-0.7695	-1.3943	-0.0067	0.66943	0.10917	1.30642
5	4	2	0.45349	-0.3762	-0.6185	-1.7157	-0.461	-1.4453	0.09464	1.55283
6	5	2	1.01492	0.94089	-0.9599	-0.8685	0.30052	0.5173	0.11447	1.38334
7	6	2	1.78265	1.70528	2.07552	0.06083	0.429	0.10902	-0.7164	-1.1645
8	7	2	0.69062	1.27761	1.15377	1.44955	-0.4663	0.24885	-1.323	-1.1454
9	8	2	0.74954	0.27013	1.05844	-0.1876	0.6692	-0.0232	1.09015	1.91198
10	9	2	1.21618	-0.2882	0.81639	1.84743	0.18484	-1.1622	-1.9981	-2.3211
11	10	2	1.20421	-0.487	-0.2274	0.34252	0.62865	1.46621	0.13736	0.13426