回帰分析と相関係数

経営では需要予測に回帰を大活用します

2025年5月7日

学習院大学経済学部教授
白田　由香利

©白田由香利、学習院大学経済学部 2024

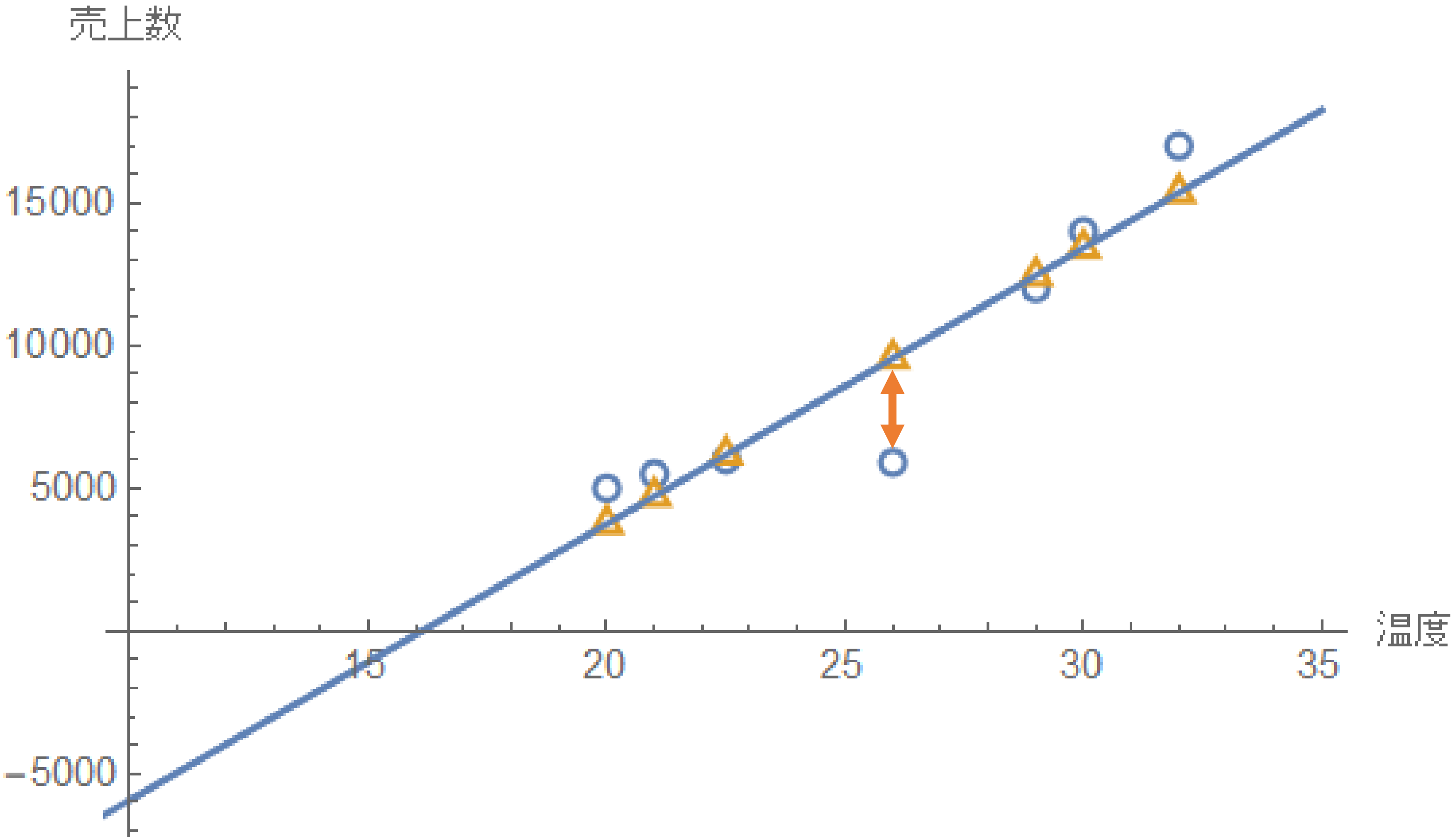
線形回帰

従来からある最もシンプルな直線でモデル化する回帰です

©白田由香利、学習院大学経済学部 2024

野球場でのビールの売上数の予測

- 観測値
- 線形回帰による予測値
- 線形単回帰の回帰モデルは直線
- $Y_i = a + bx_i$

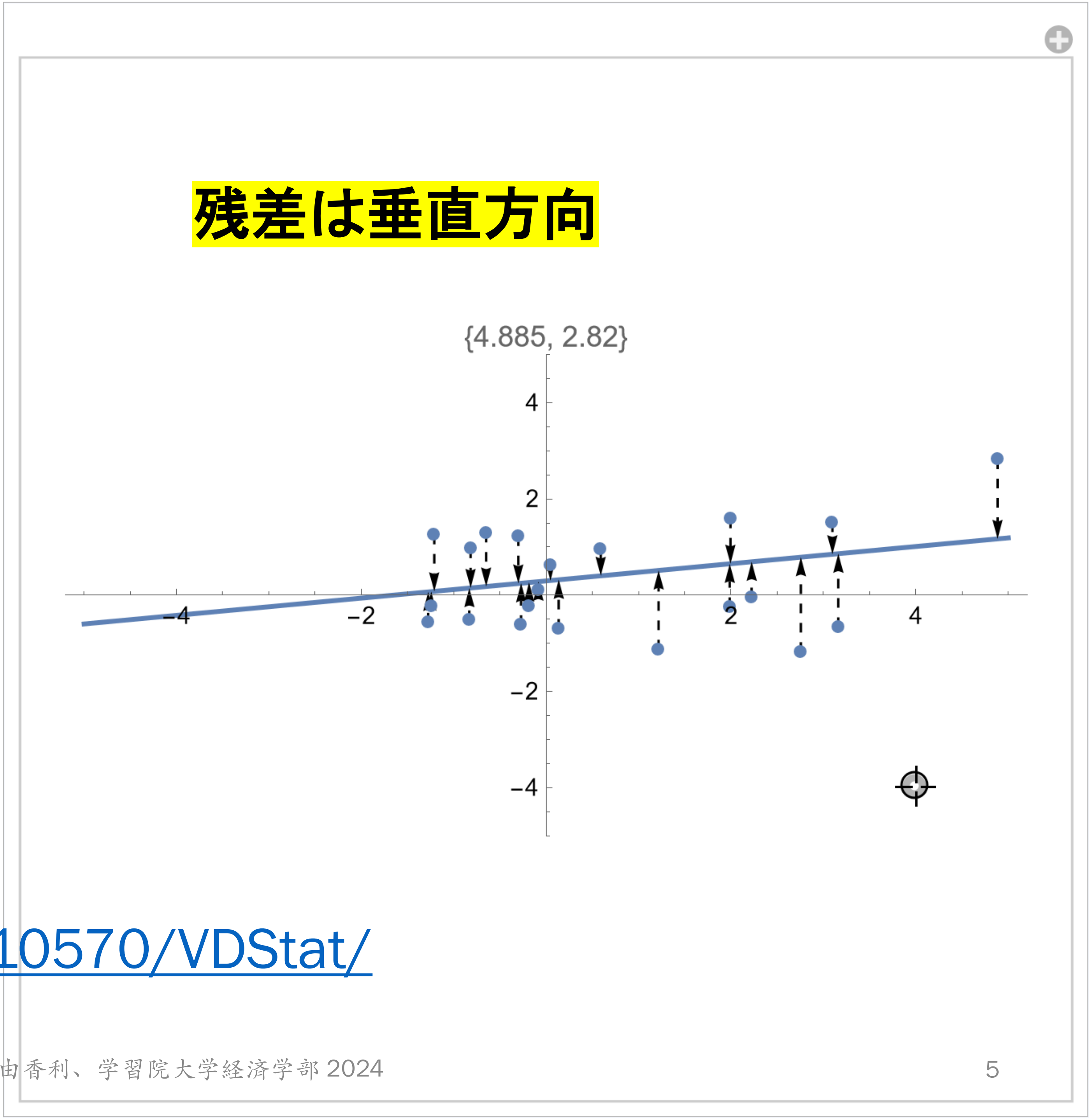


白田グラフィクス教材

- www-cc.gakushuin.ac.jp/~20010570/VDStat/
- Wolfram CDF player(無料）で動かします
- 大学のPCのブラウザ上で動かせます
- 希望者は，自宅PCの場合，CDF playerをインストールしてください

回帰分析の視覚化
最後の1点を動かして回帰式の変化を見る

- 外れ値があると回帰モデルが大きく変わってしまう



[グラフィクス教材](www-cc.gakushuin.ac.jp/~20010570/VDStat/)
www-cc.gakushuin.ac.jp/~20010570/VDStat/

©白田由香利、学習院大学経済学部 2024

残差：観測値ー予測値

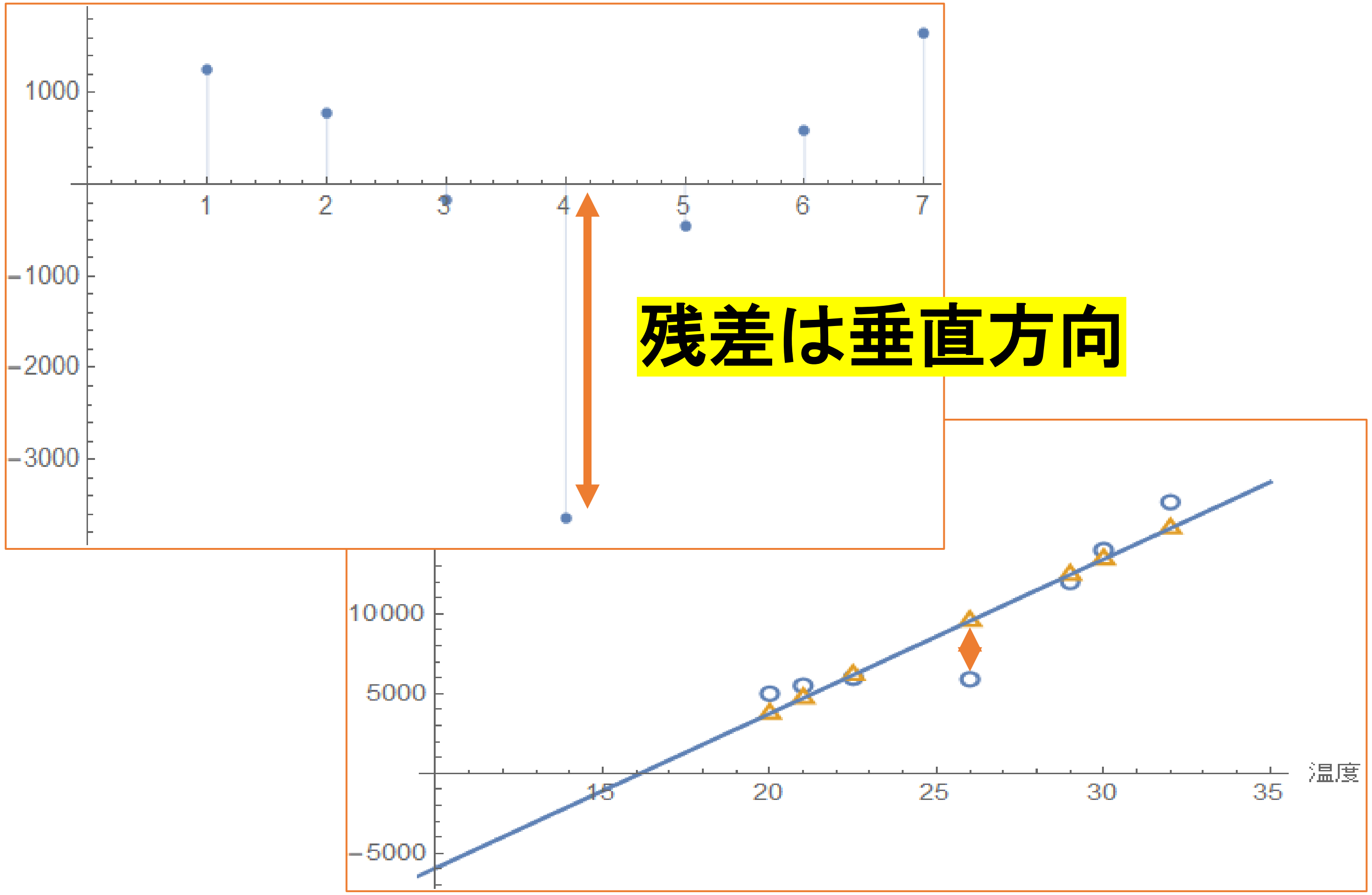
$$y_i - Y_i = y_i - (a + bx_i)$$

- 観測値 y_i
- 予測値 Y_i
- 推定された回帰式

$$Y_i = a + bx_i$$

- 残差　$y_i - (a + bx_i)$

$$\begin{pmatrix} 5000 - a - 20\,b \\ 5500 - a - 21\,b \\ 6000 - a - 22.5\,b \\ 5900 - a - 26\,b \\ 12\,000 - a - 29\,b \\ 14\,000 - a - 30\,b \\ 17\,000 - a - 32\,b \end{pmatrix}$$



最小二乗法
残差の平方和を最小にす a, bを求める

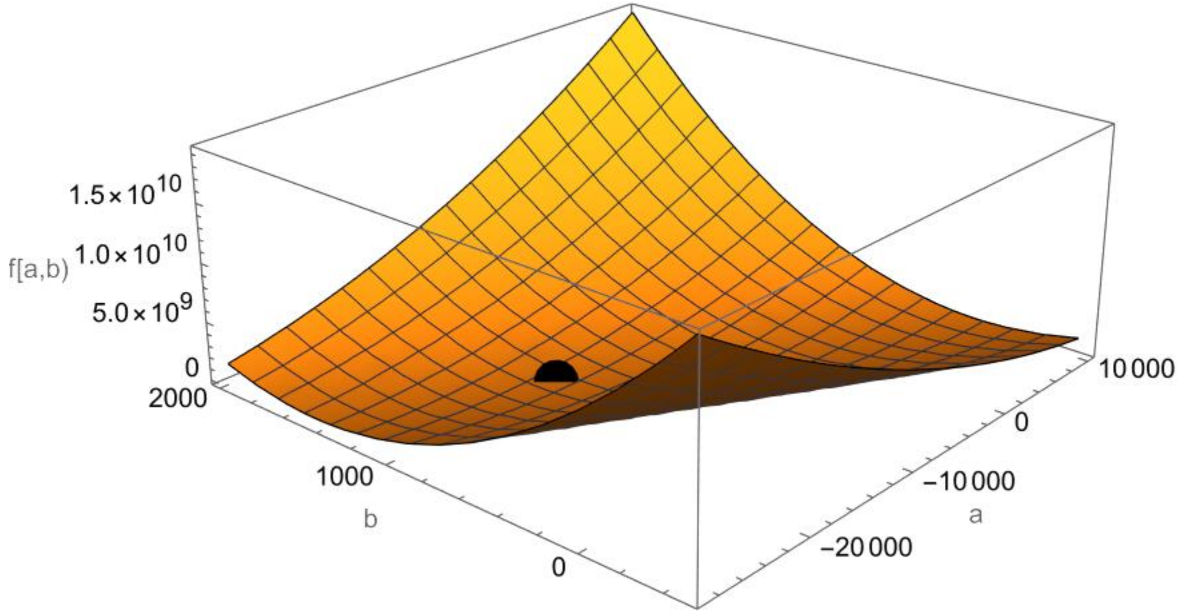
f(a,b)=

$(17000-a-32\,b)^2+(14000-a-30\,b)^2+(12000-a-29\,b)^2+(5900-a-26\,b)^2+(6000-a-22.5\,b)^2+(5500-a-21\,b)^2+(5000-a-20\,b)^2$

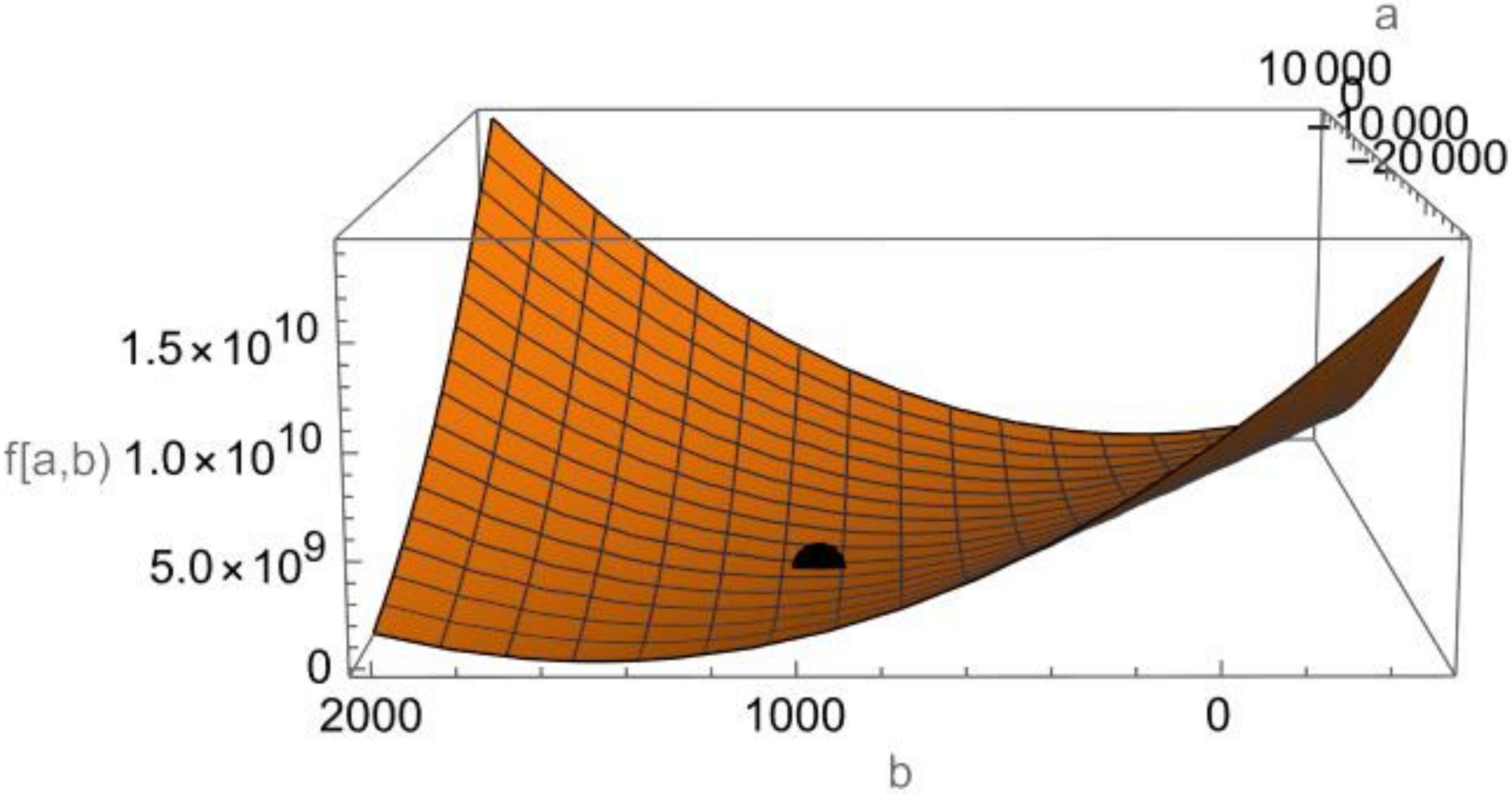
- $\frac{\partial f}{\partial a} = 0, \quad \frac{\partial f}{\partial b} = 0$

偏微分して
この2元連立方程式を解く

- {a= -15593., b= 967.04}
- **回帰式**$Y_i = -15593 + 967.04x_i$

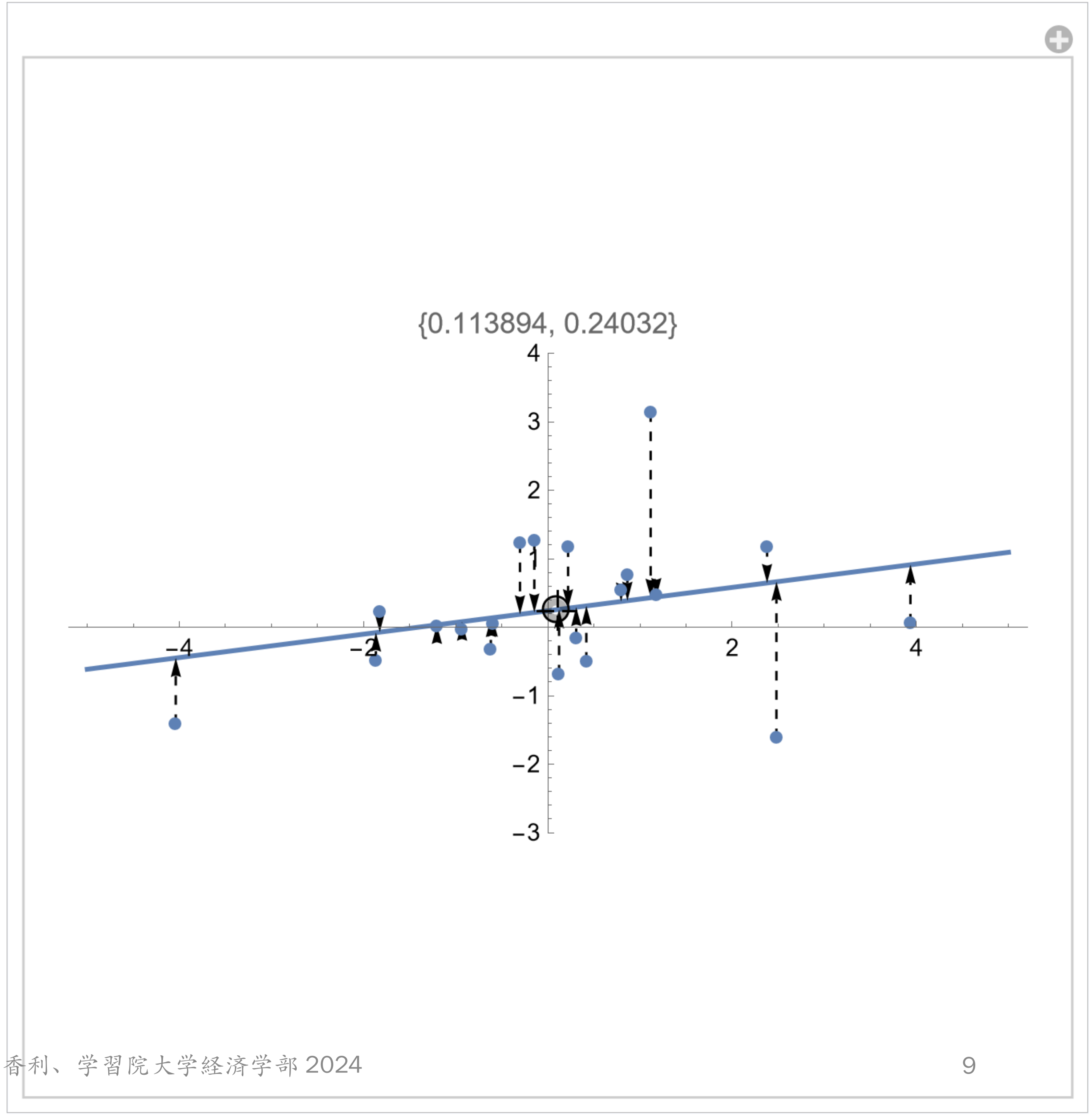


最小二乗法
残差の平方和を最小にす a, bを求める



重心を原点に

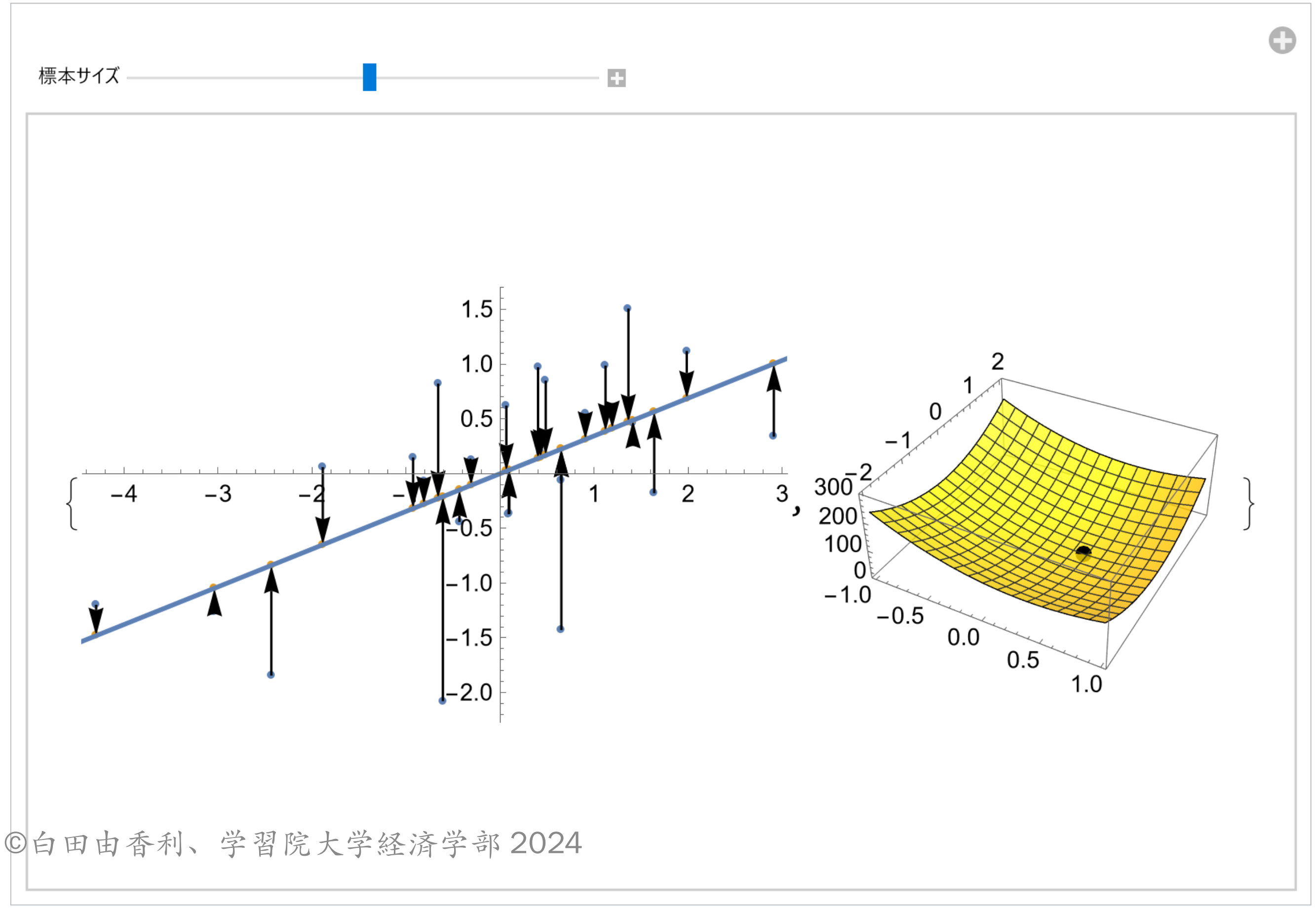
- 重心（平均値）を計算
- 平均値分を各データから引き算する
- 重心が原点に移動する



グラフィクス教材

www-cc.gakushuin.ac.jp/~20010570/VDStat/

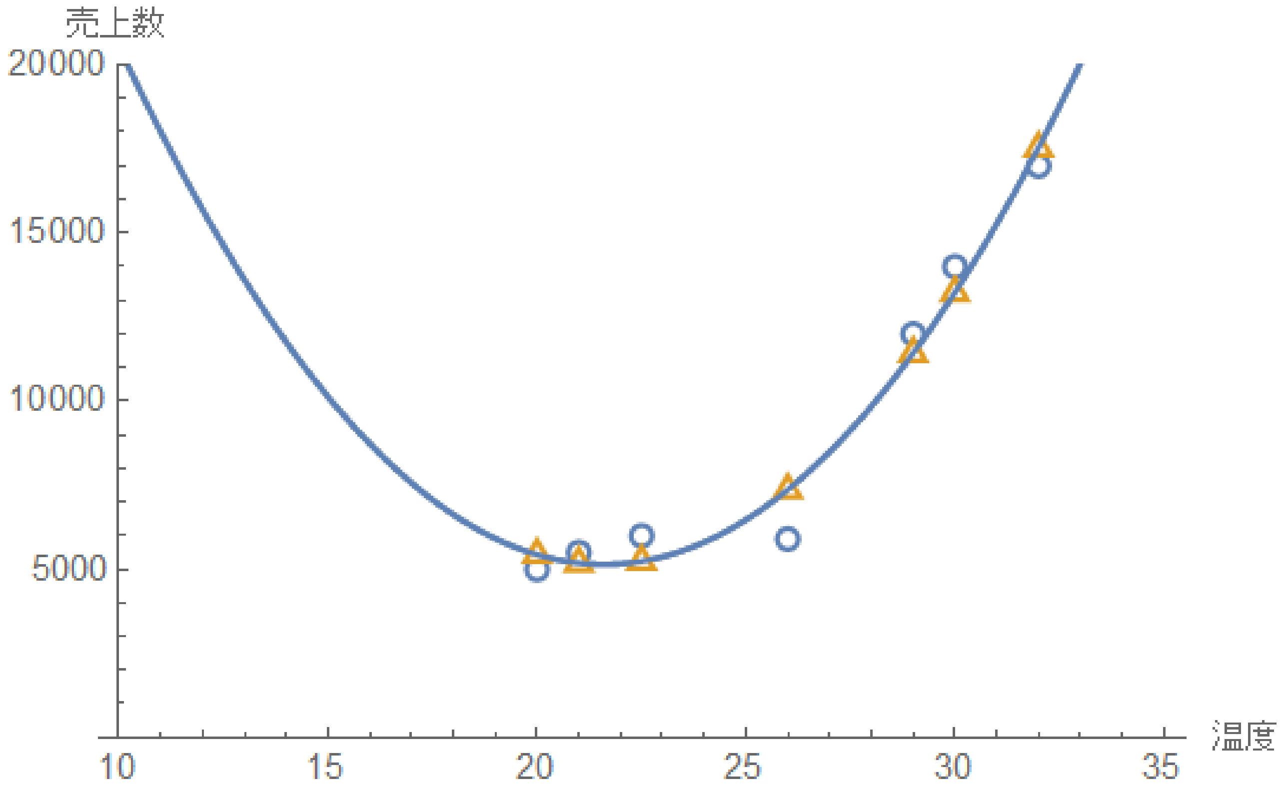
- 最小二乗法
  残差の平方和を最小にす a, bを求める



©白田由香利、学習院大学経済学部 2024

線形回帰と2次関数による回帰

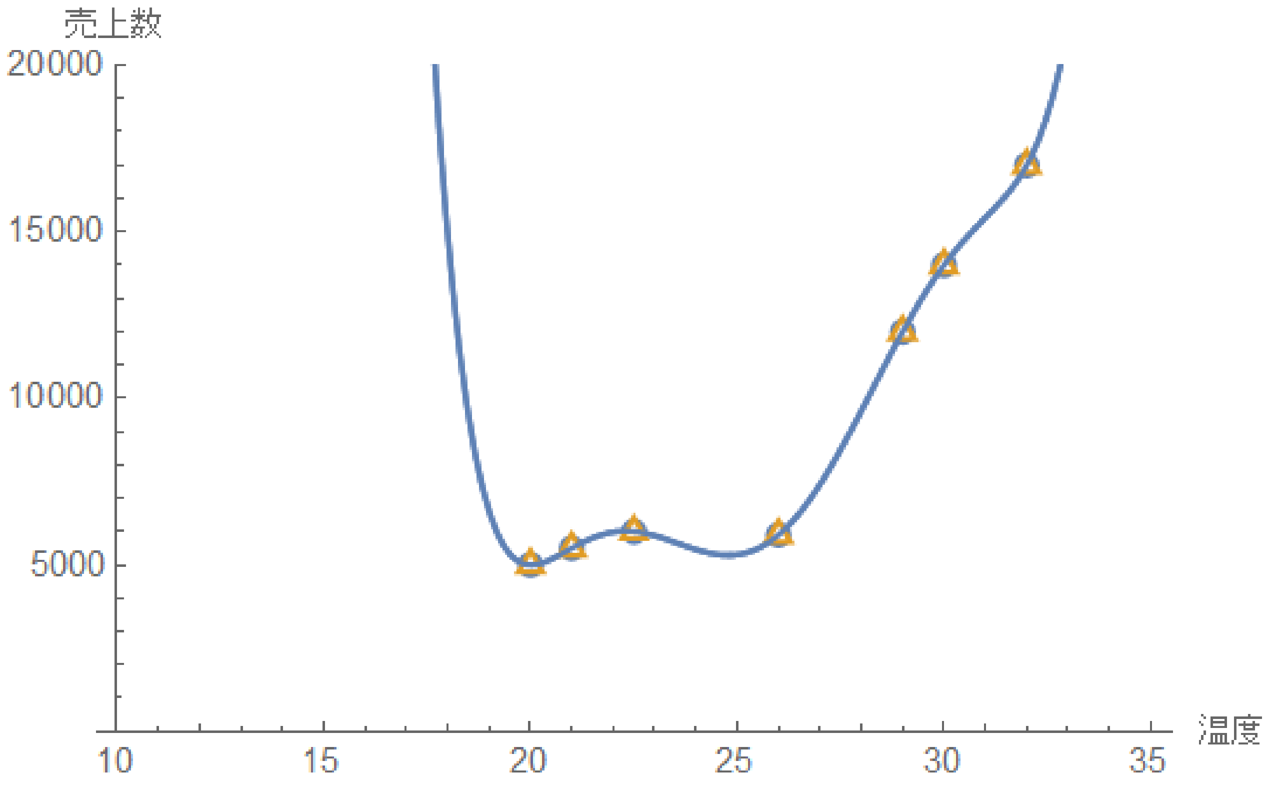
- 線形回帰とは1次式でモデル化すること
- 2次式で回帰モデルを作ってみる
- $Y = 58501 - 4942x + 114x^2$



データ7個の場合、6次式で完全一致

$Y = 6.00857*10^7 -$
$1.44597*10^7 x$
$+1.43874*10^6 x^2$
$-75746.8 x^3$
$+2225.41 x^4$
$-34.5954 x^5$
$+0.222359 x^6$

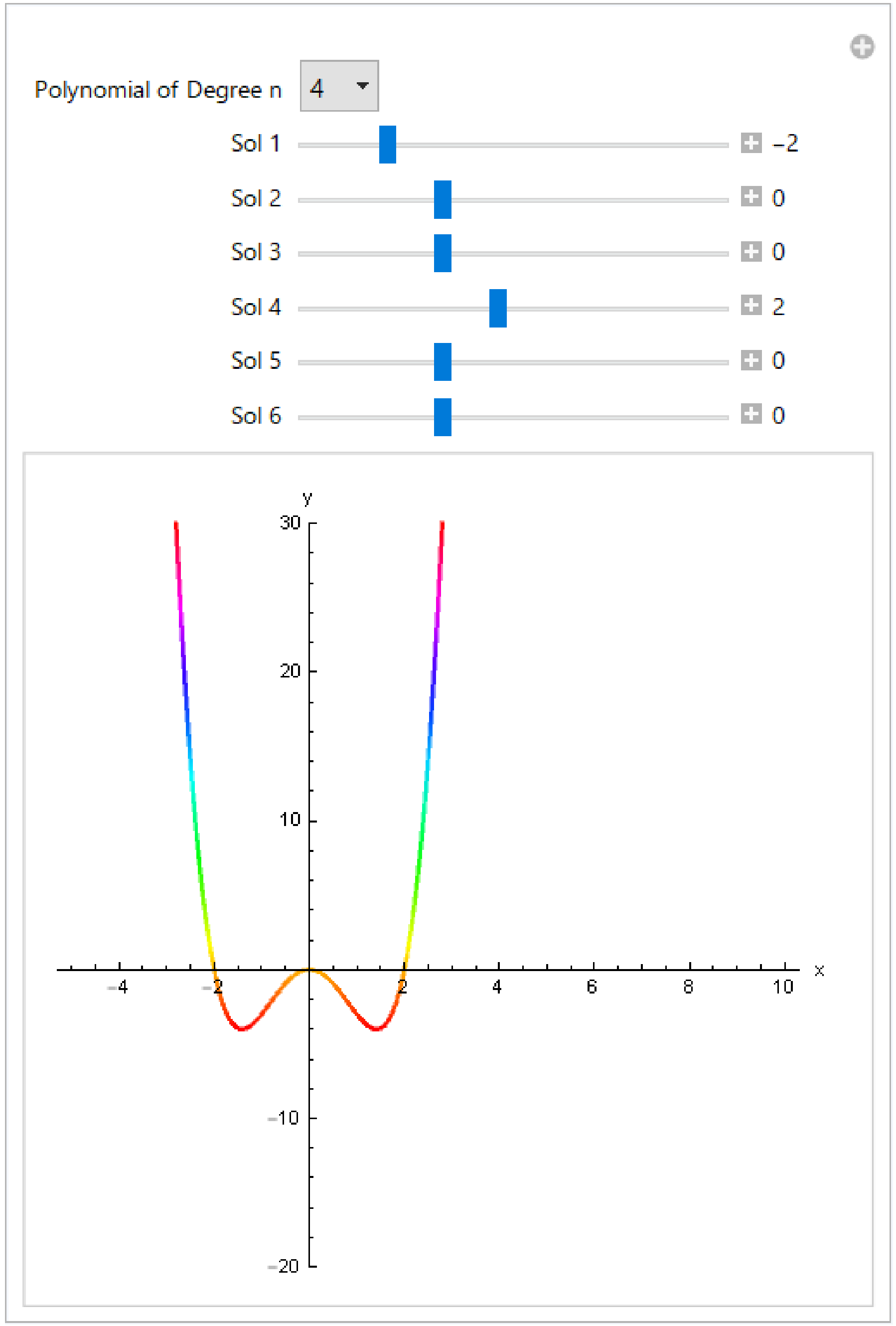
しかし、
オーバーフィットで、
未知のｘに対して
良い予測ができなくなる



高次関数形状をみる

- https://shirotaabc.sakura.ne.jp/usefulMath/ABC/index.html

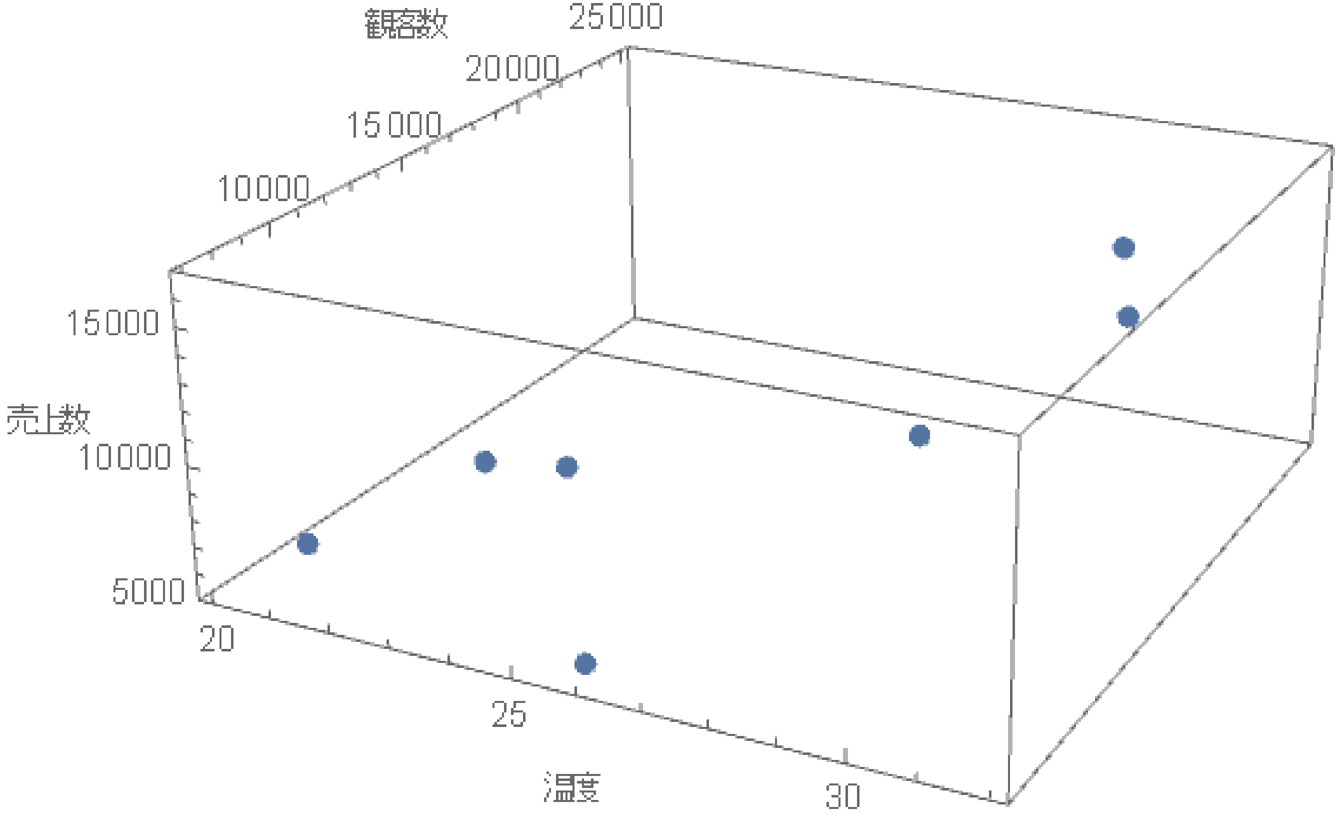
  上記の一番下にビデオがあるので、一度DLしてから再生せよ。
- 学習院大学　白田 グラフィクス教材サイト (sakura.ne.jp)



重回帰分析　２説明変数の場合

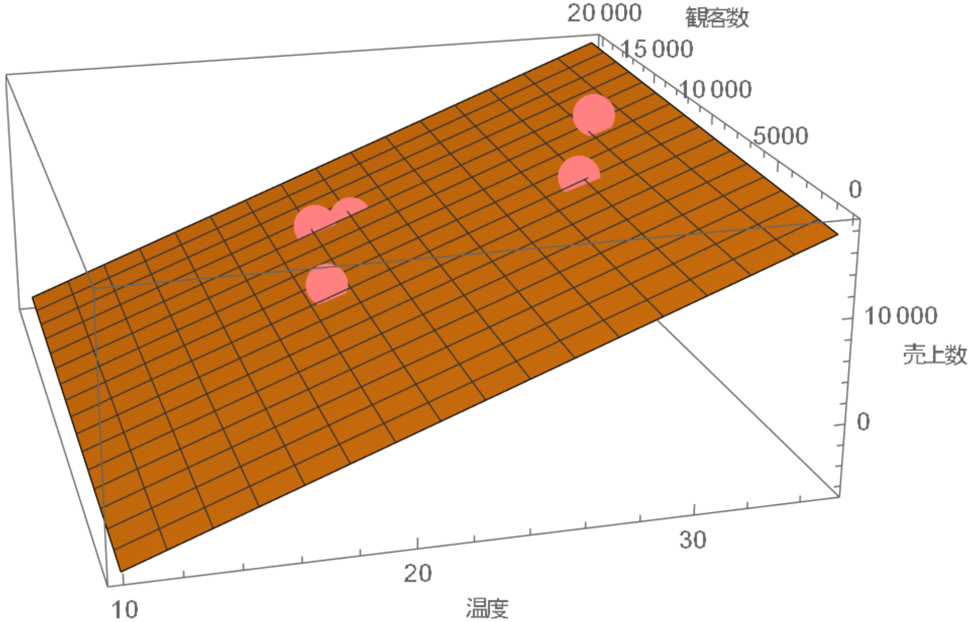
- 温度、観客数

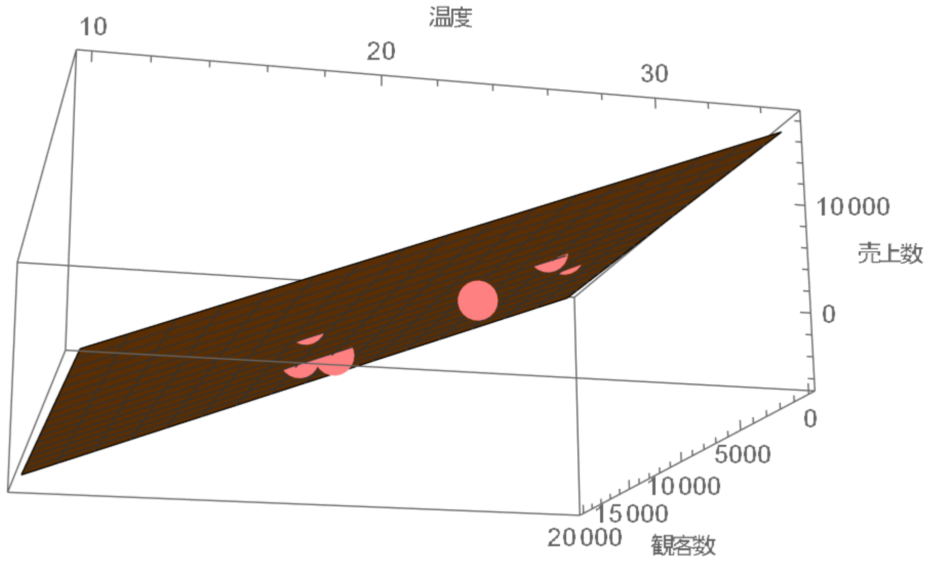
- 7個の観測値
- これを通るような
  平面を見つける



重回帰分析　２説明変数の場合

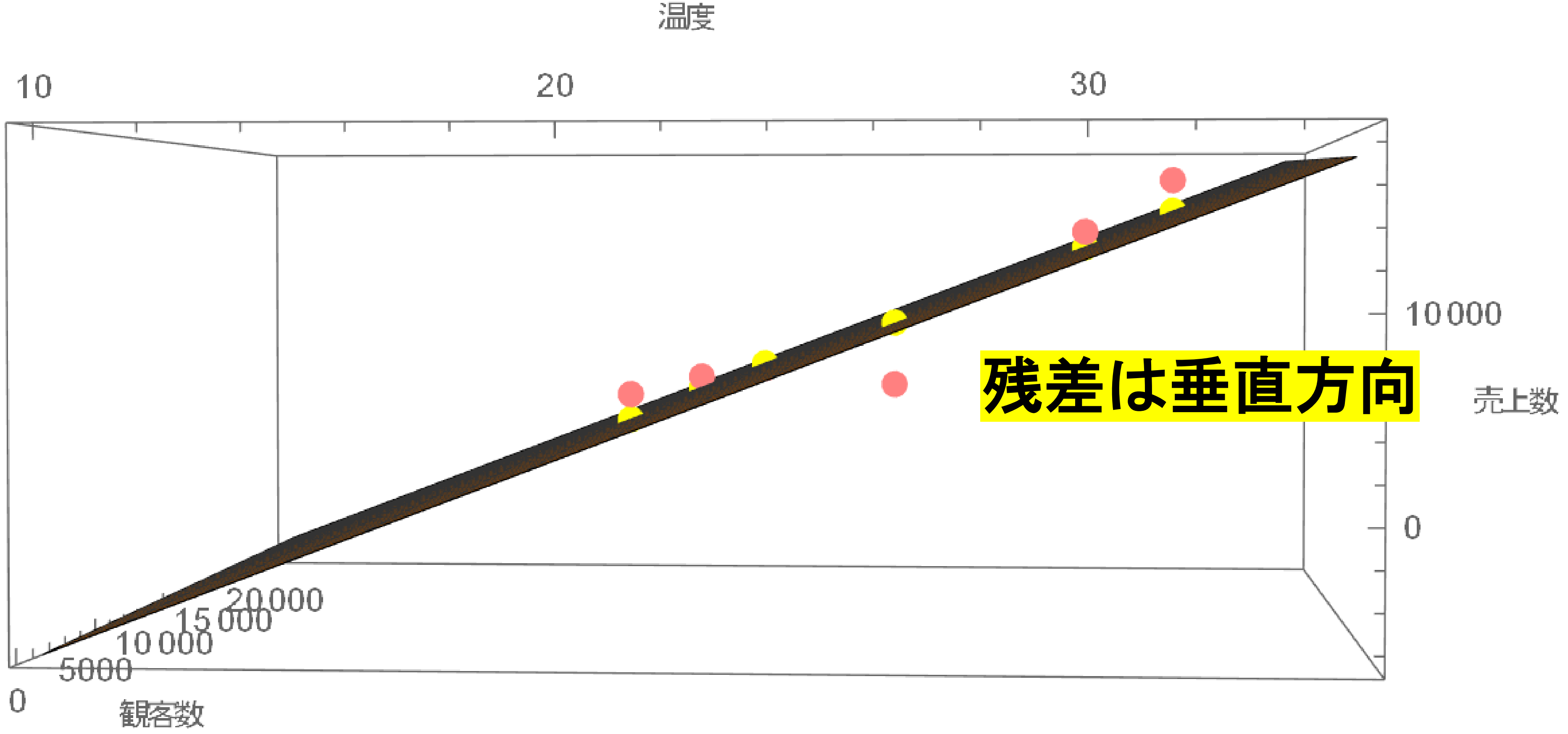
- 温度 x
- 観客数 y
- $f(x, y) = -16120 + 956x + 0.06y$





重回帰分析　２説明変数の場合
観測値と予測値

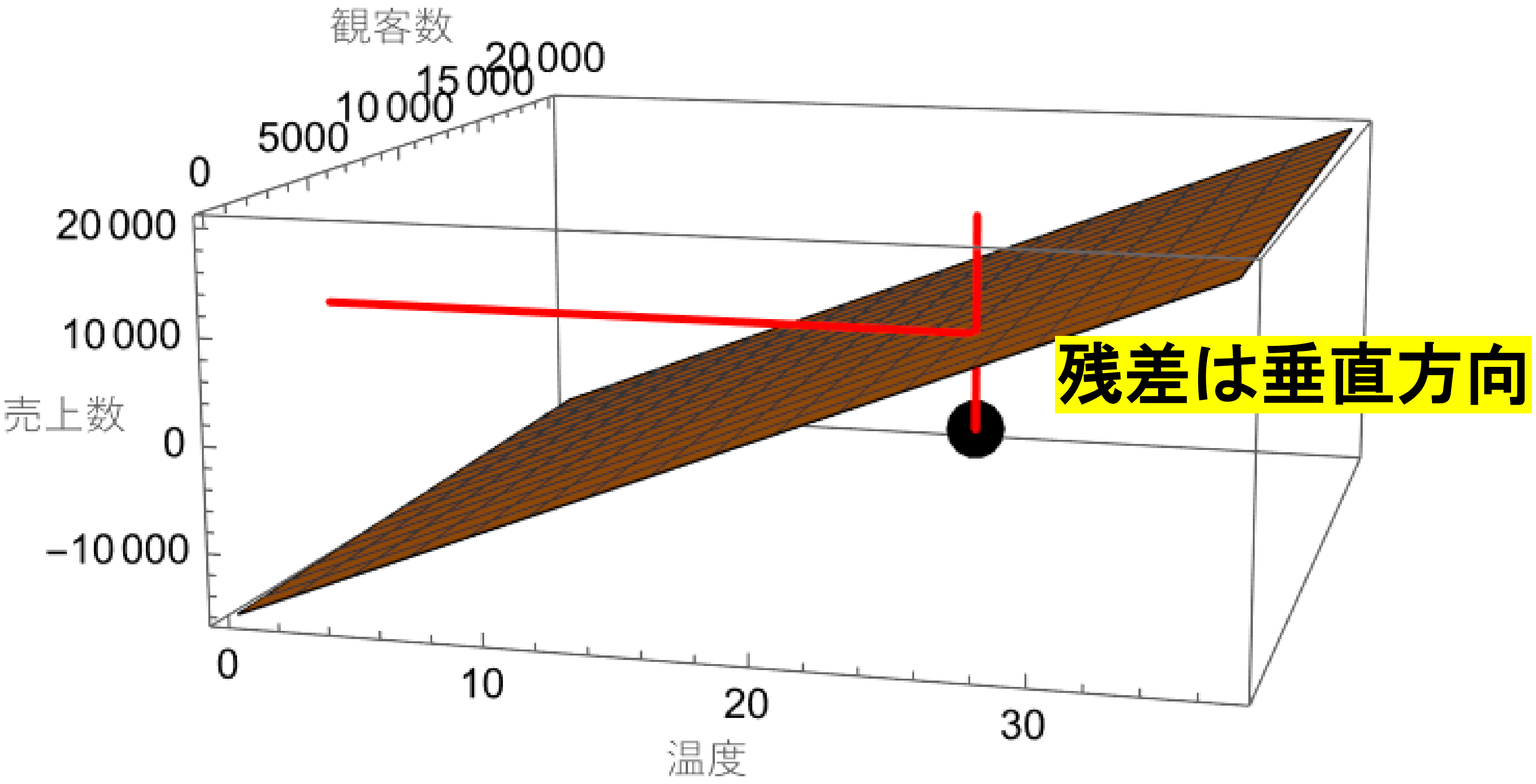
- $f(x, y) = -16120 + 956x + 0.06y$



©白田由香利、学習院大学経済学部 2024

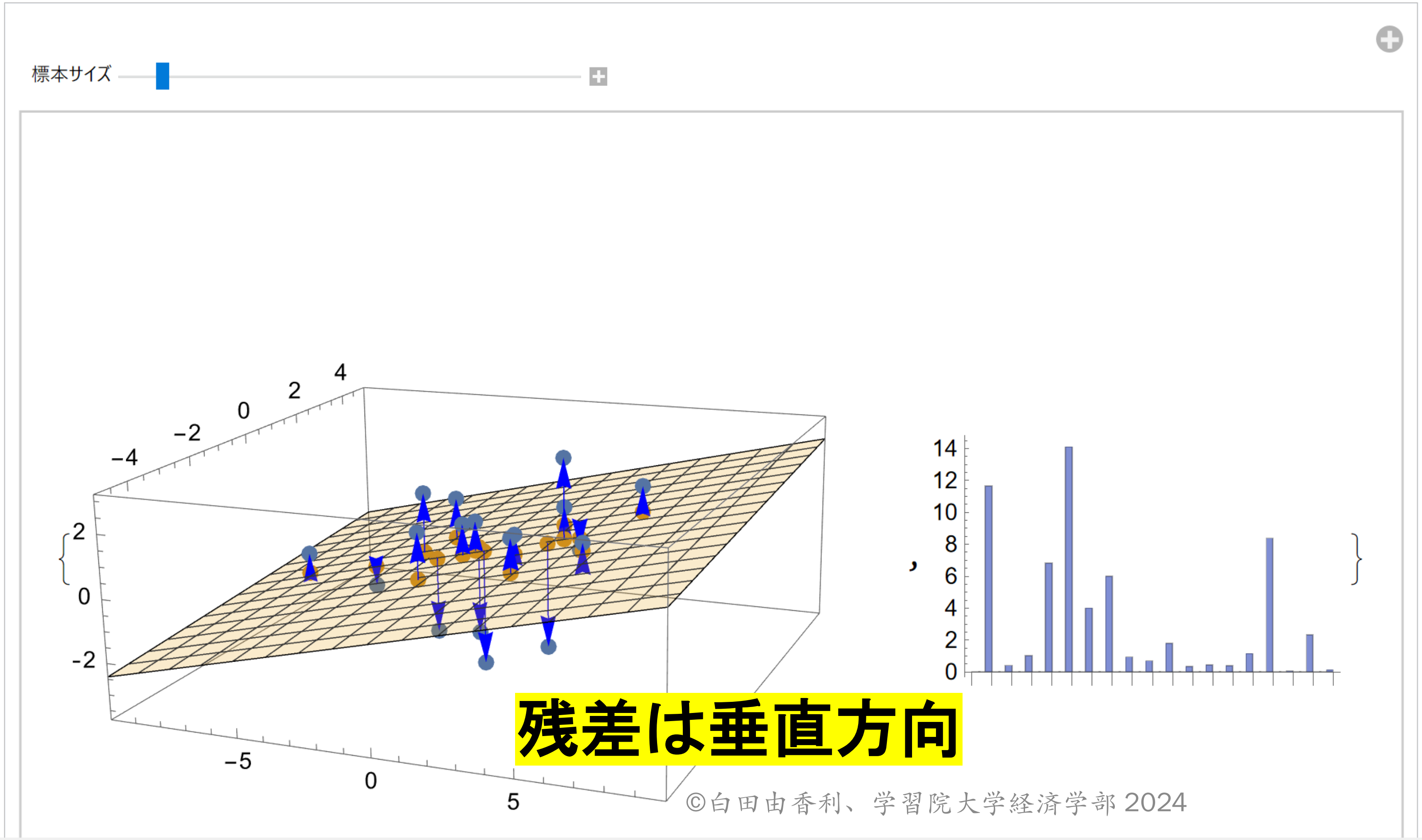
重回帰分析　２説明変数の場合 温度26度，観客5000人で予測

- 需要予測：9032個

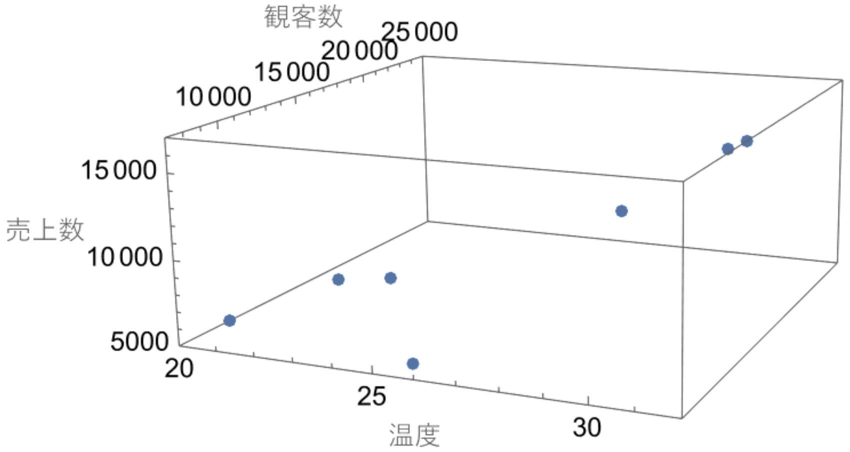


グラフィクス教材
www-cc.gakushuin.ac.jp/~20010570/VDStat/

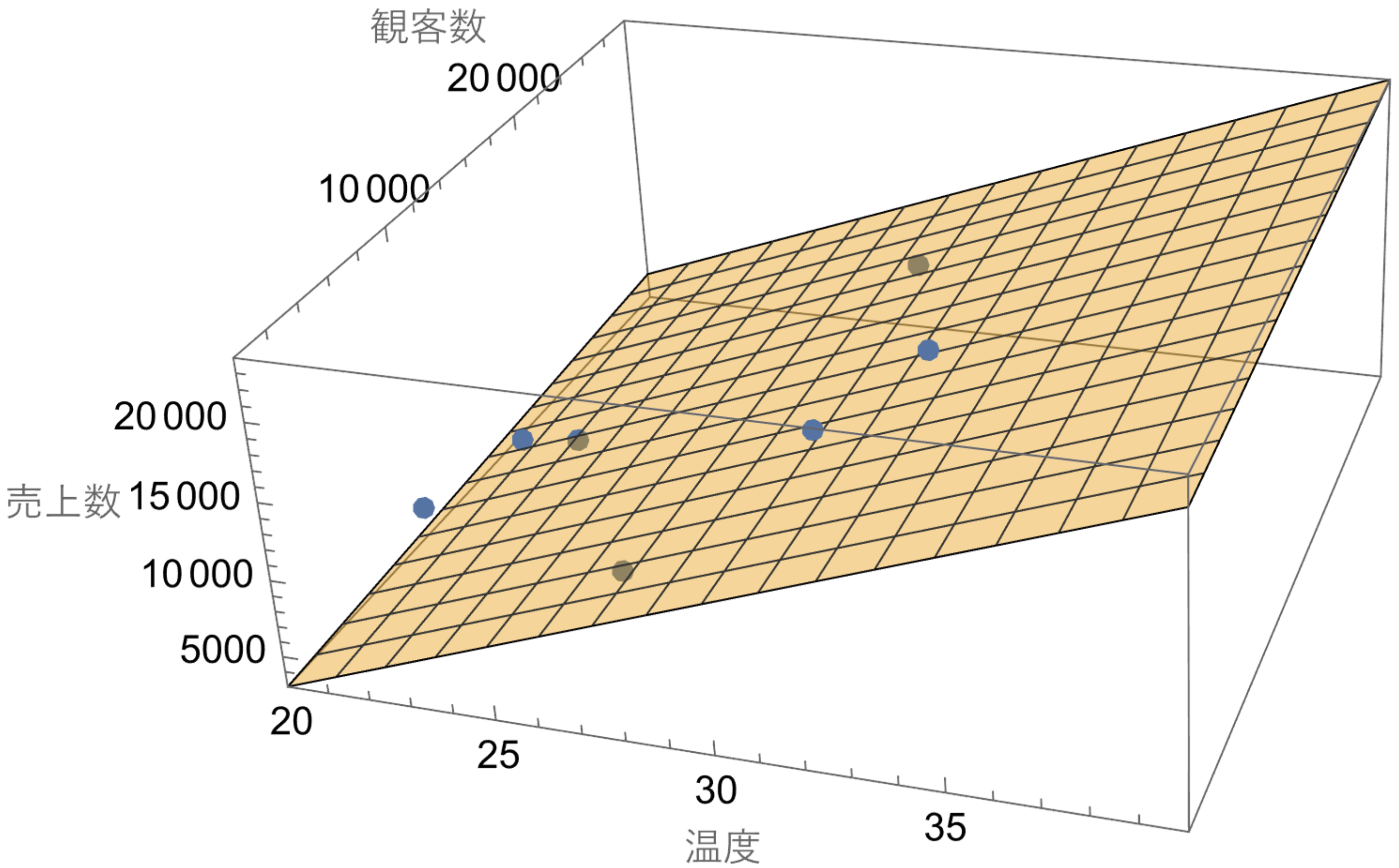
残差平方和を最小にするように回帰平面を作る



線形重回帰



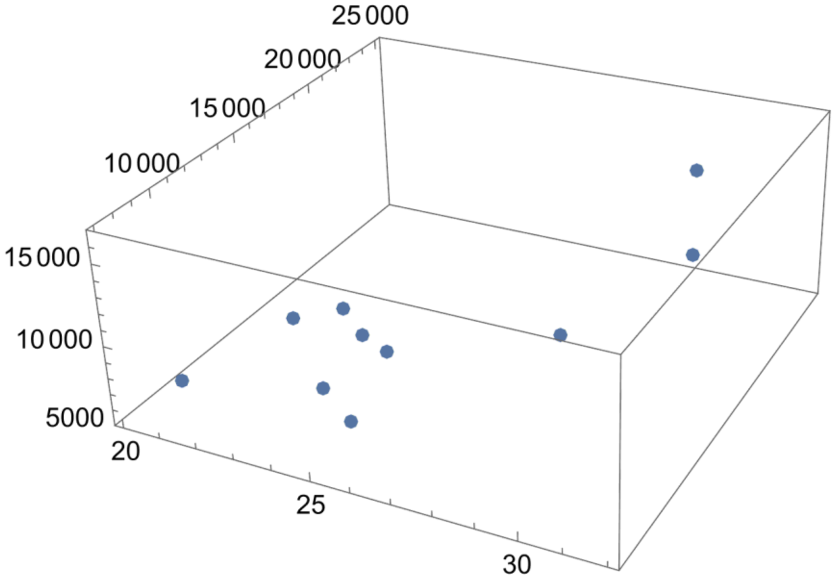
観客数
25 000
20 000
15 000
10 000
15 000
売上数
10 000
5000
20
25
30
温度

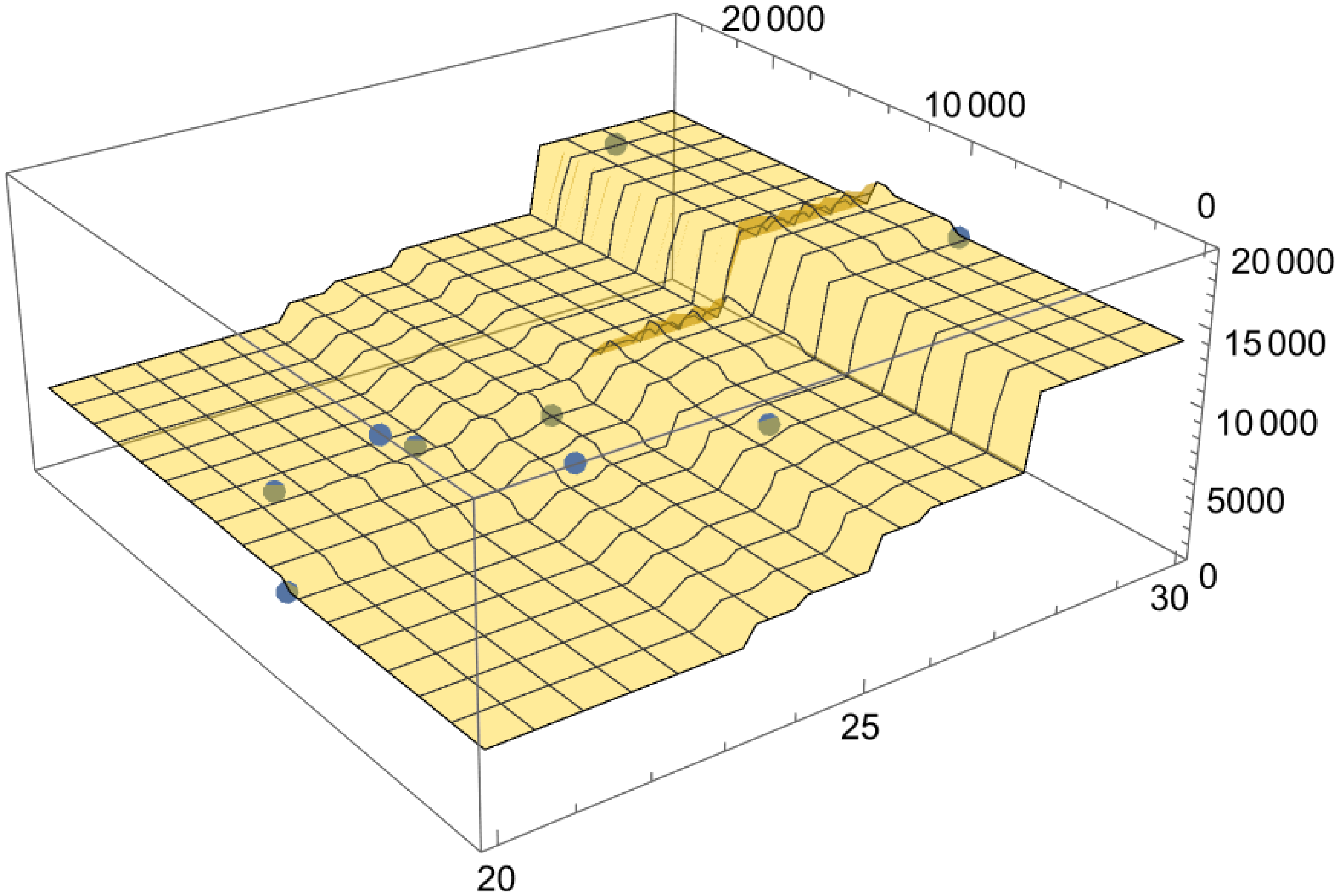


観客数
20 000
10 000
20 000
売上数
15 000
10 000
5000
20
25
30
35
温度

AIによる回帰　XGBOOST

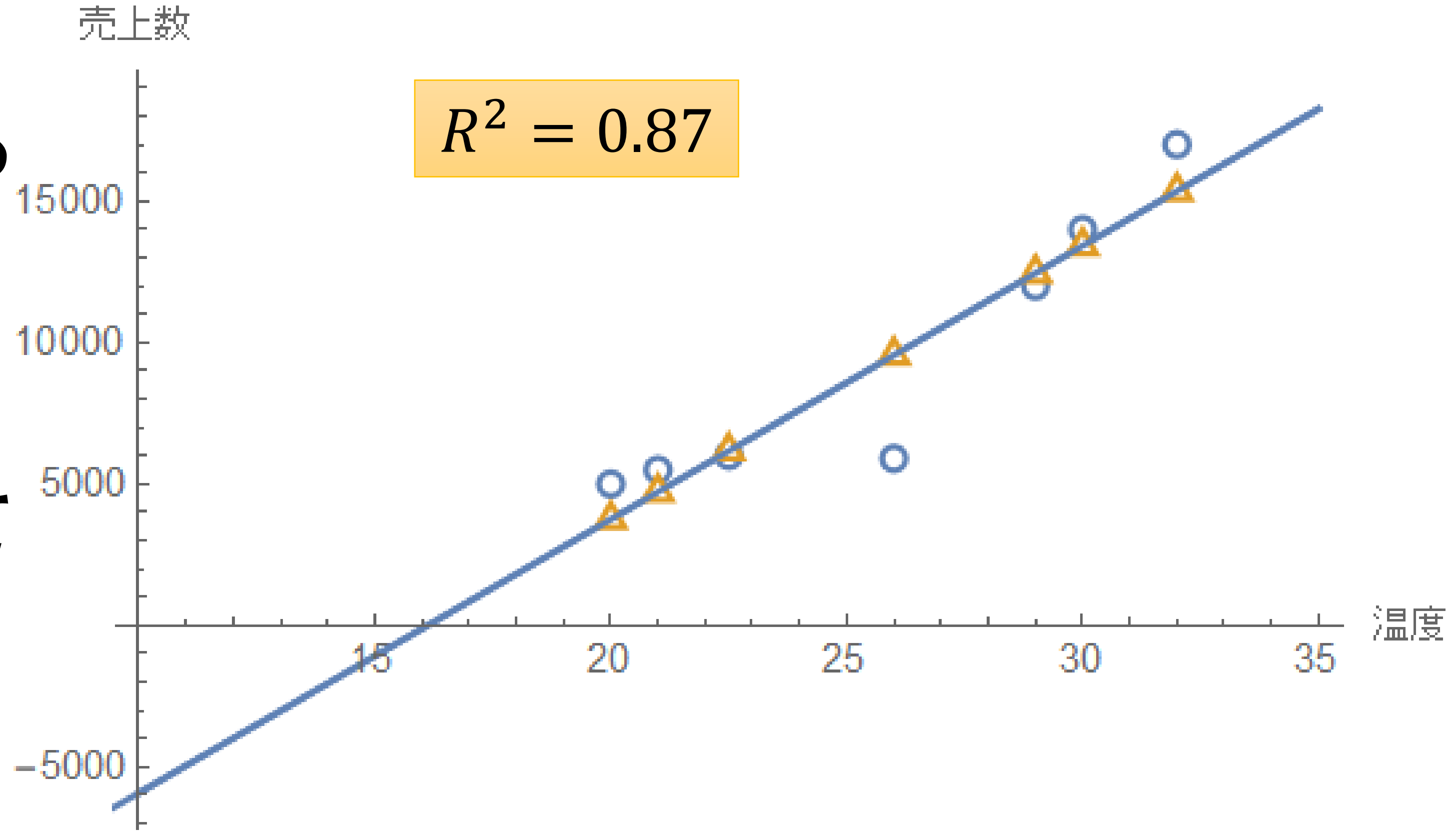
-





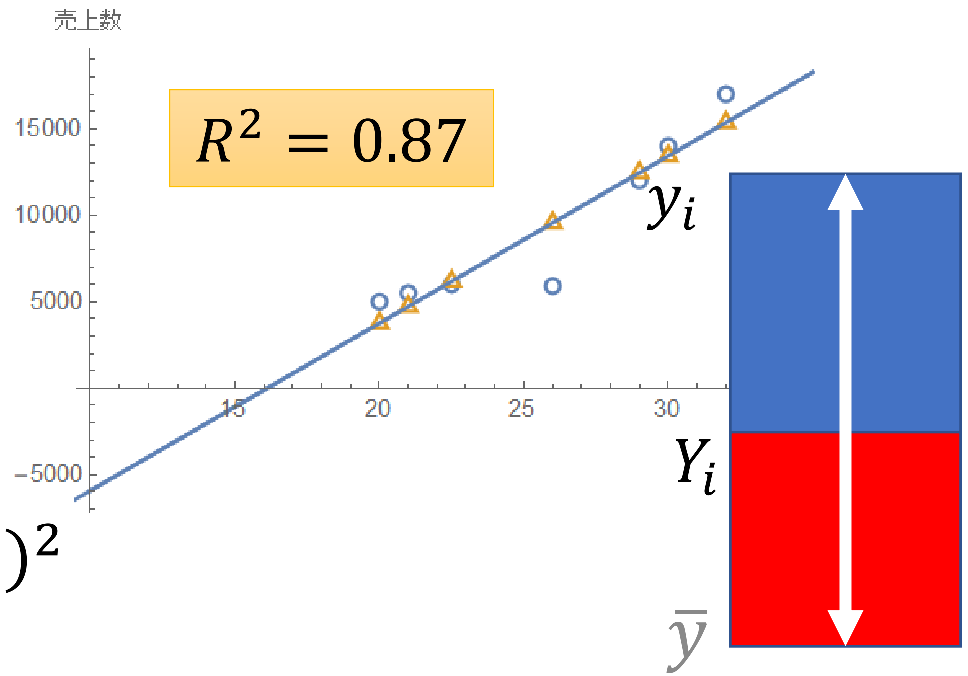
決定係数
フィッティングの良さ

- １に近いほど良い
- ０から１の値をとる
- 回帰分析では
  0.7以上はほしい
- 例えば、0.3では
  回帰が意味をなしていない



決定係数
フィッティングの良さ

- $R^2 = \dfrac{\text{(予測値で説明された変動)}}{\text{(偏差の平方和)}}$
- 恒等式
$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(Y_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - Y_i)^2$



偏差の平方和＝予測値で説明された変動＋予測値で説明されなかった変動(残差の平方和)

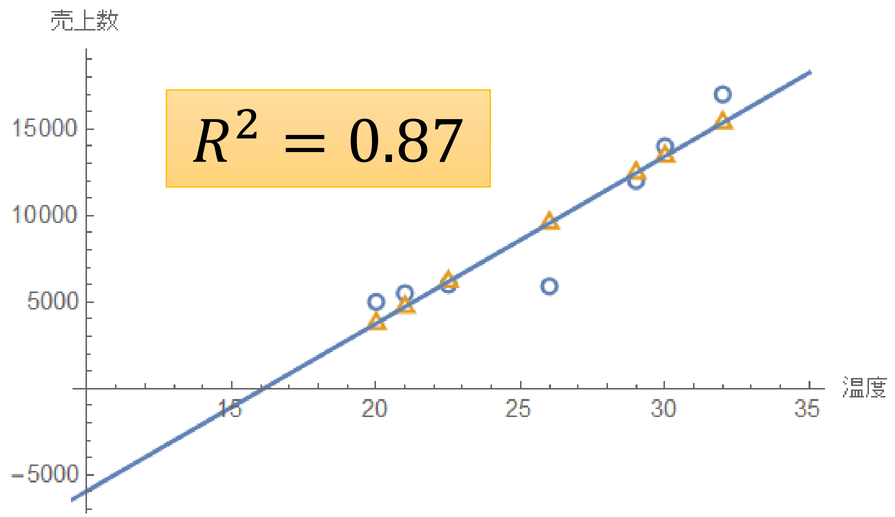
偏差とは観測値の平均$\bar{y}$からのずれ

偏差は観測値の平均からのずれ
残差は観測値と予測値のずれ

観測値の平均と、予測値の平均はいつでも一致する

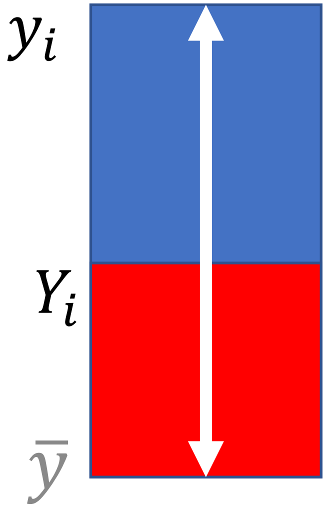
©白田由香利、学習院大学経済学部 2024

決定係数
フィッティングの良さ



- $R^2 = \dfrac{(\text{予測値で説明された変動})}{(\text{偏差の平方和})}$
- **恒等式**
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(Y_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - Y_i)^2$$

偏差の平方和＝予測値で説明された変動＋予測値で説明されなかった変動(残差の平方和)

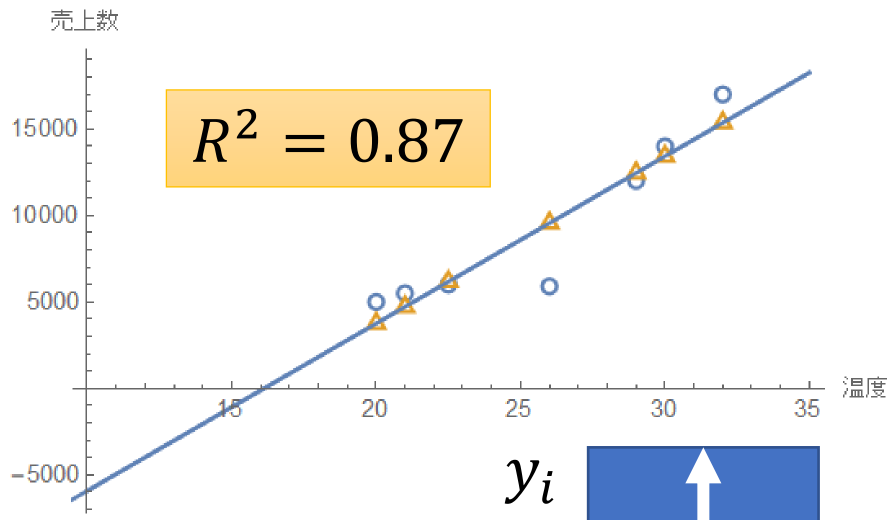


偏差とは平均$\bar{y}$からのずれ

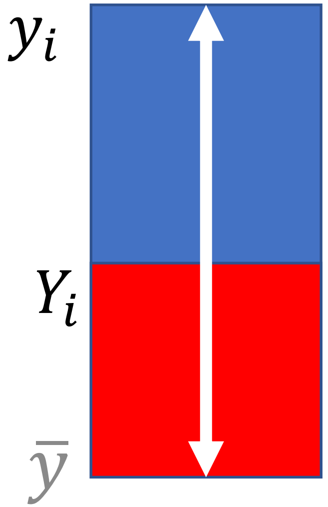
観測値の平均と、予測値の平均はいつでも一致する

©白田由香利、学習院大学経済学部 2024

決定係数
フィッティングの良さ



- $R^2 = \frac{\text{(予測値で説明された変動)}}{\text{(偏差の平方和)}}$
- 恒等式の両辺を偏差の平方和で割る
- $$\frac{(\sum_{i=1}^{n}(y_i-\bar{y})^2)}{(\sum_{i=1}^{n}(y_i-\bar{y})^2)} = \frac{\sum_{i=1}^{n}(Y_i-\bar{y})^2}{(\sum_{i=1}^{n}(y_i-\bar{y})^2)} + \frac{\sum_{i=1}^{n}(y_i-Y_i)^2}{(\sum_{i=1}^{n}(y_i-\bar{y})^2)}$$



$$1 = \boxed{\frac{\sum_{i=1}^{n}(Y_i-\bar{y})^2}{(\sum_{i=1}^{n}(y_i-\bar{y})^2)}} + \frac{\sum_{i=1}^{n}(y_i-Y_i)^2}{(\sum_{i=1}^{n}(y_i-\bar{y})^2)}$$

$\sum_{i=1}^{n}(y_i - Y_i)^2$残差の項が大きくなれば,
決定係数は小さくなる

©白田由香利、学習院大学経済学部 2024

回帰分析の結果　R2≅0.85

- 変動から自分で計算可能
- 自由度は
 １＋３＝４
- 残差の変動の自由度
n-2＝5ー2＝３

データの取得と変換　クエリと接続　並べ替えとフィルター　データ ツ

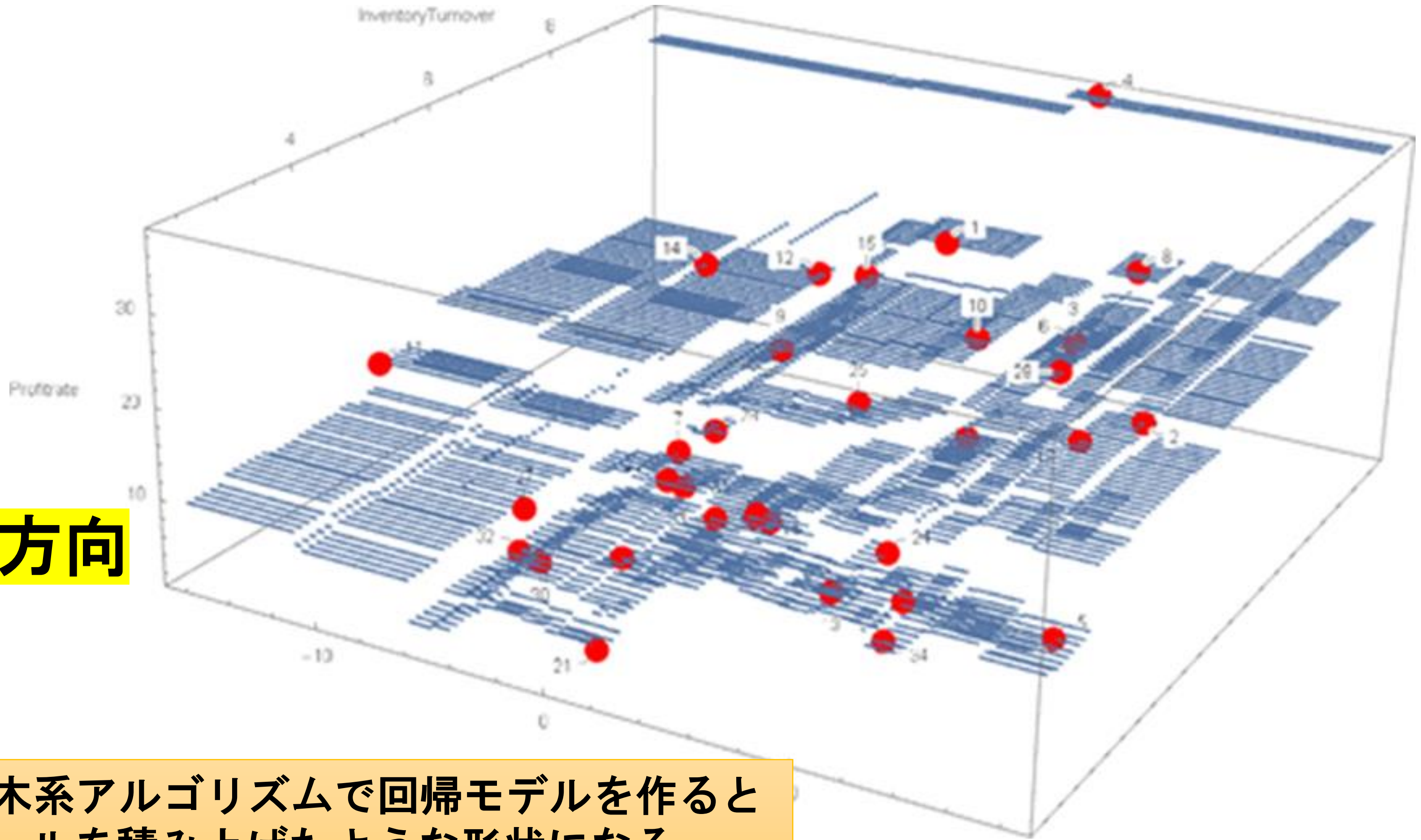
YEAR　=1-C13/C14

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.924472							
5	重決定 R2	0.854648		=1-C13/C14					
6	補正 R2	0.806197							
7	標準誤差	963.7924							
8	観測数	5							
9									
10	分散分析表								
11		自由度	変動	分散	観測された分散比	有意 F			
12	回帰	1	16385313	16385312.66	17.63955981	0.024633			
13	残差	3	2786687	928895.7796					
14	合計	4	19172000						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 5.0%	上限 5.0%
17	切片	-10361.6	4573.184328	-2.265737617	0.108350337	-24915.5	4192.278	-10673	-10050.3
18	温度	709.2595	168.8734134	4.199947596	0.024632892	171.8829	1246.69	697.7613	720.7577
19									

機械学習の回帰分析
回帰モデルは複雑形状

- 垂直軸：予測値
- 赤丸：観測値
- 予測値と観測値のずれ→R^2で評価
  線形回帰と同じR^2

残差は垂直方向

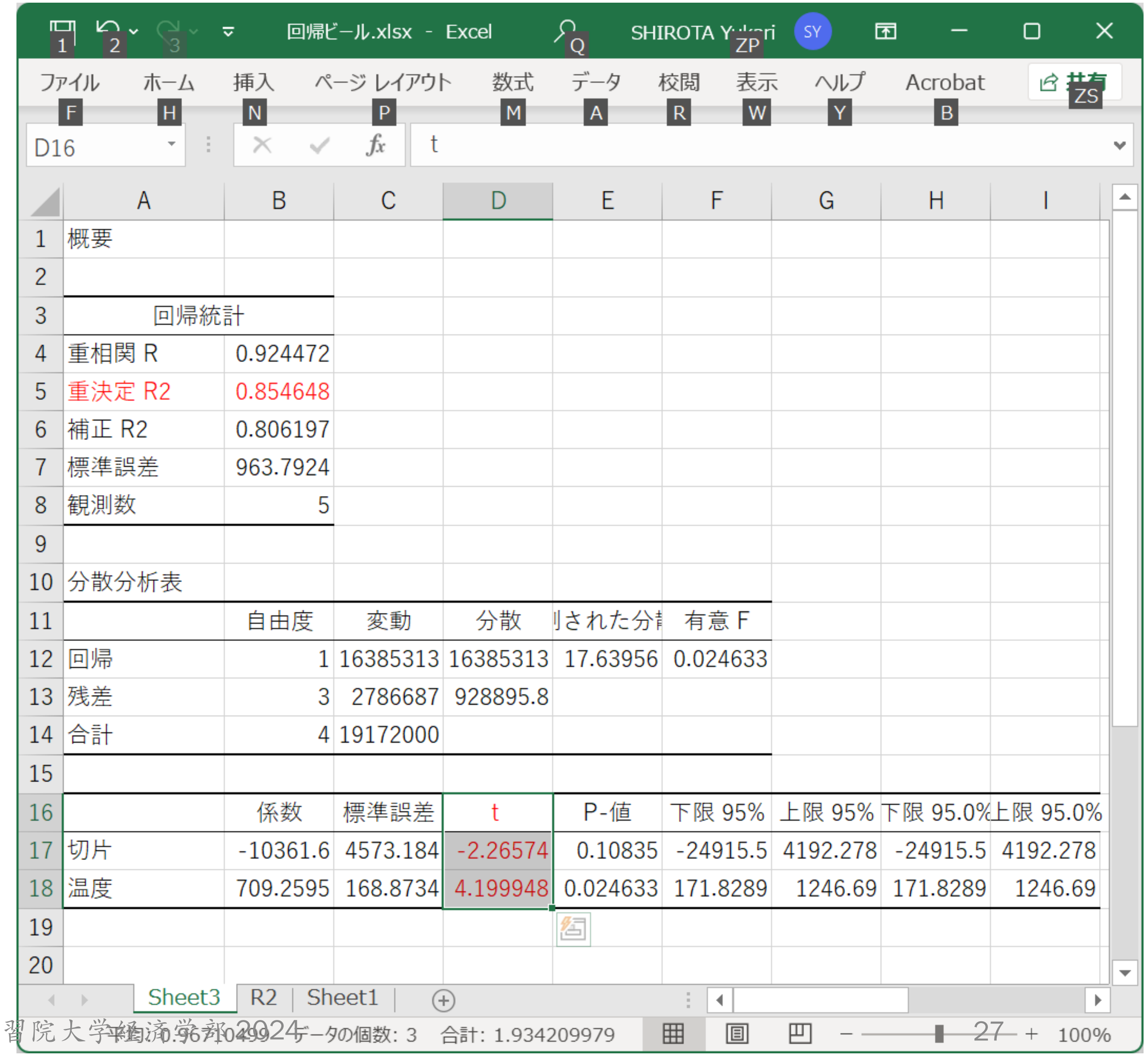


XGBOOSTのような樹木系アルゴリズムで回帰モデルを作ると
この図のような段ボールを積み上げたような形状になる

回帰分析の結果の t値とは

統計的検定を先に学ぶこと

- 回帰の切片と温度の傾き
- 温度xは需要yに本当に影響を与えているのか，t検定で調べている
- たまたま観測日7日をこのように選択したら，a, bが
  -10361.6, 709と求められた
- $Y_i = a + bx_i$
- 違う観測日にしたら，異なるa, bの値となる．
- では何度も繰り返してやってみる→分布ができる

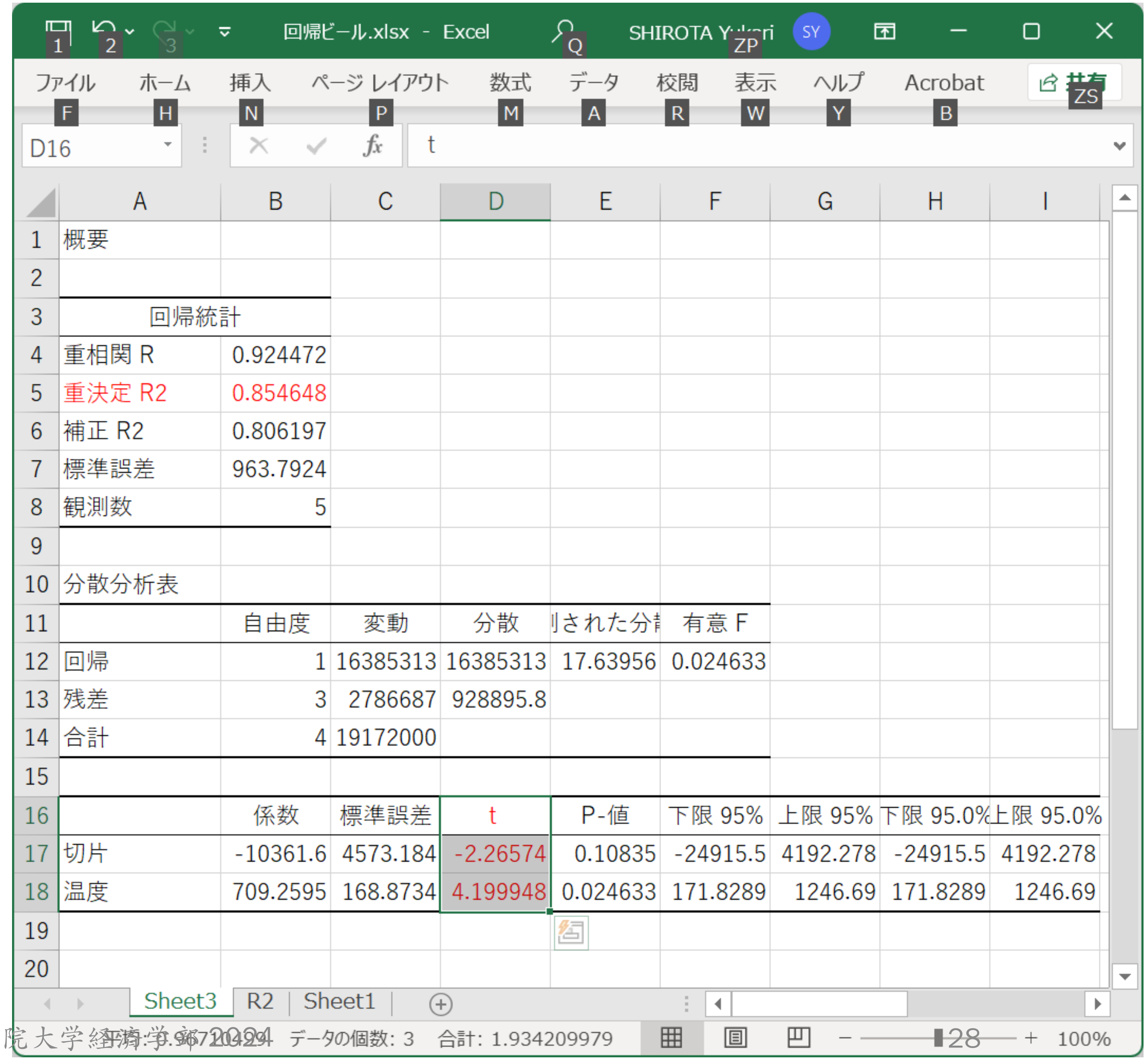


	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.924472							
5	重決定 R2	0.854648							
6	補正 R2	0.806197							
7	標準誤差	963.7924							
8	観測数	5							
9									
10	分散分析表								
11		自由度	変動	分散	された分	有意 F			
12	回帰	1	16385313	16385313	17.63956	0.024633			
13	残差	3	2786687	928895.8					
14	合計	4	19172000						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	-10361.6	4573.184	-2.26574	0.10835	-24915.5	4192.278	-24915.5	4192.278
18	温度	709.2595	168.8734	4.199948	0.024633	171.8289	1246.69	171.8289	1246.69
19									
20									

©白田由香利、学習院大学経済学部 2024

回帰分析の結果の t値

- $Y_i = a + bx_i$
- 何度も繰り返し違うデータで回帰したとする．{bi}, i=1…n
- 無限大に近い回数をすると，{bi} はどのような分布になるのか？ → 正規分布
- $b \sim N\left(B, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}\right)$
- 真の回帰係数Bと誤差項の分散σ^2が分かっていた場合，正規分布

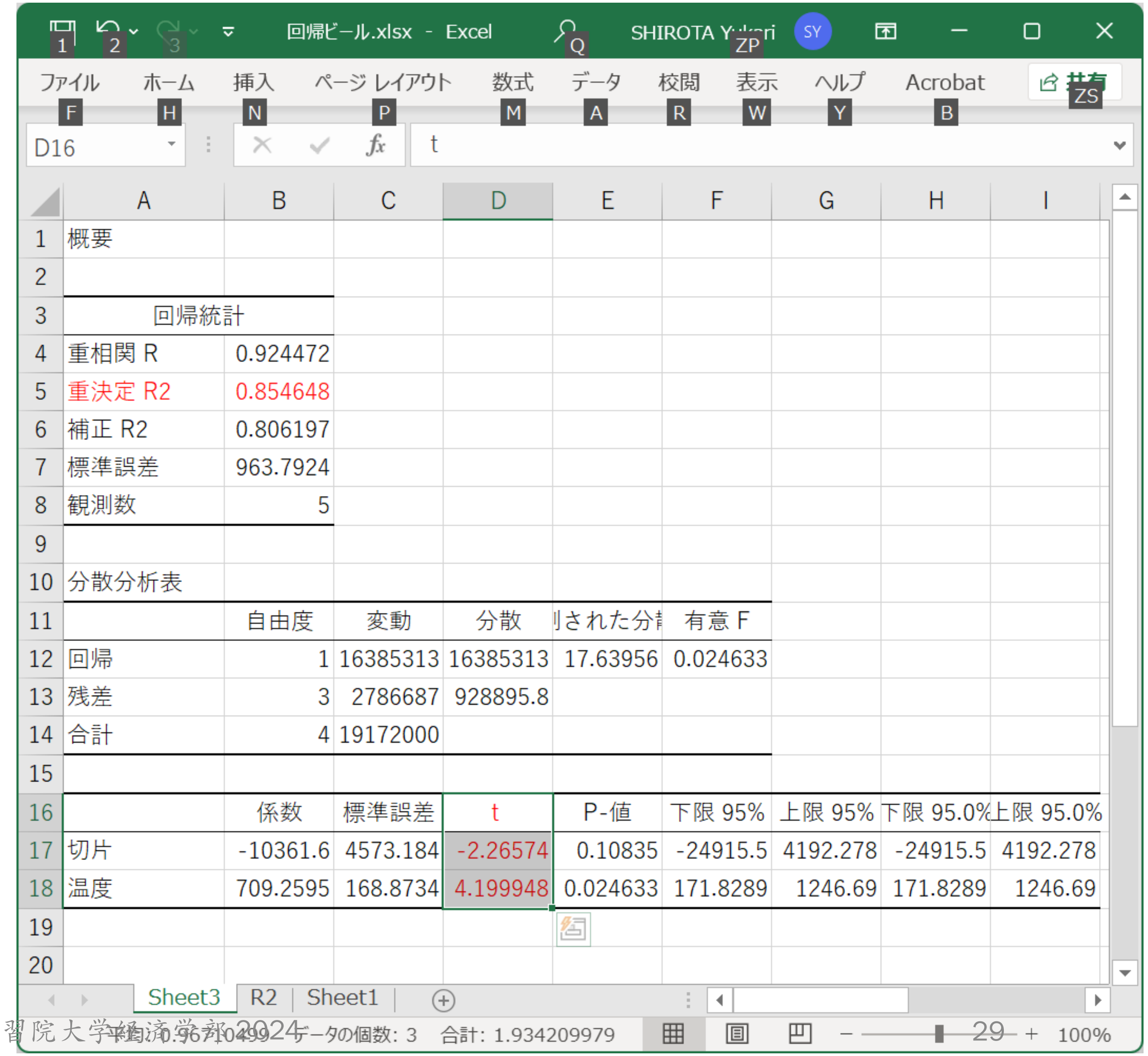


	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.924472							
5	重決定 R2	0.854648							
6	補正 R2	0.806197							
7	標準誤差	963.7924							
8	観測数	5							
9									
10	分散分析表								
11		自由度	変動	分散	された分	有意 F			
12	回帰	1	16385313	16385313	17.63956	0.024633			
13	残差	3	2786687	928895.8					
14	合計	4	19172000						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	-10361.6	4573.184	-2.26574	0.10835	-24915.5	4192.278	-24915.5	4192.278
18	温度	709.2595	168.8734	4.199948	0.024633	171.8289	1246.69	171.8289	1246.69
19									
20									

©白田由香利、学習院大学経済学部 2024

回帰分析の結果の t値とは

- $Y_i = a + bx_i$
- $b \sim N(B, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2})$
- 真の回帰係数Bと真の誤差項の分散σ^2が分かっていた場合，正規分布
- しかし，σ^2が分かっていないので t 分布を用いてt検定する→(n-2)の自由度のt分布
- σ^2の代用として s^2 =［残差の平方和÷（n-2)］を使う
- 真のBからどれだけずれているか
- $t = \frac{b - B}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}} = \frac{b - B}{b\text{の標準誤差}}$



	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.924472							
5	重決定 R2	0.854648							
6	補正 R2	0.806197							
7	標準誤差	963.7924							
8	観測数	5							
9									
10	分散分析表								
11		自由度	変動	分散	された分	有意 F			
12	回帰	1	16385313	16385313	17.63956	0.024633			
13	残差	3	2786687	928895.8					
14	合計	4	19172000						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	-10361.6	4573.184	-2.26574	0.10835	-24915.5	4192.278	-24915.5	4192.278
18	温度	709.2595	168.8734	4.199948	0.024633	171.8289	1246.69	171.8289	1246.69
19									
20									

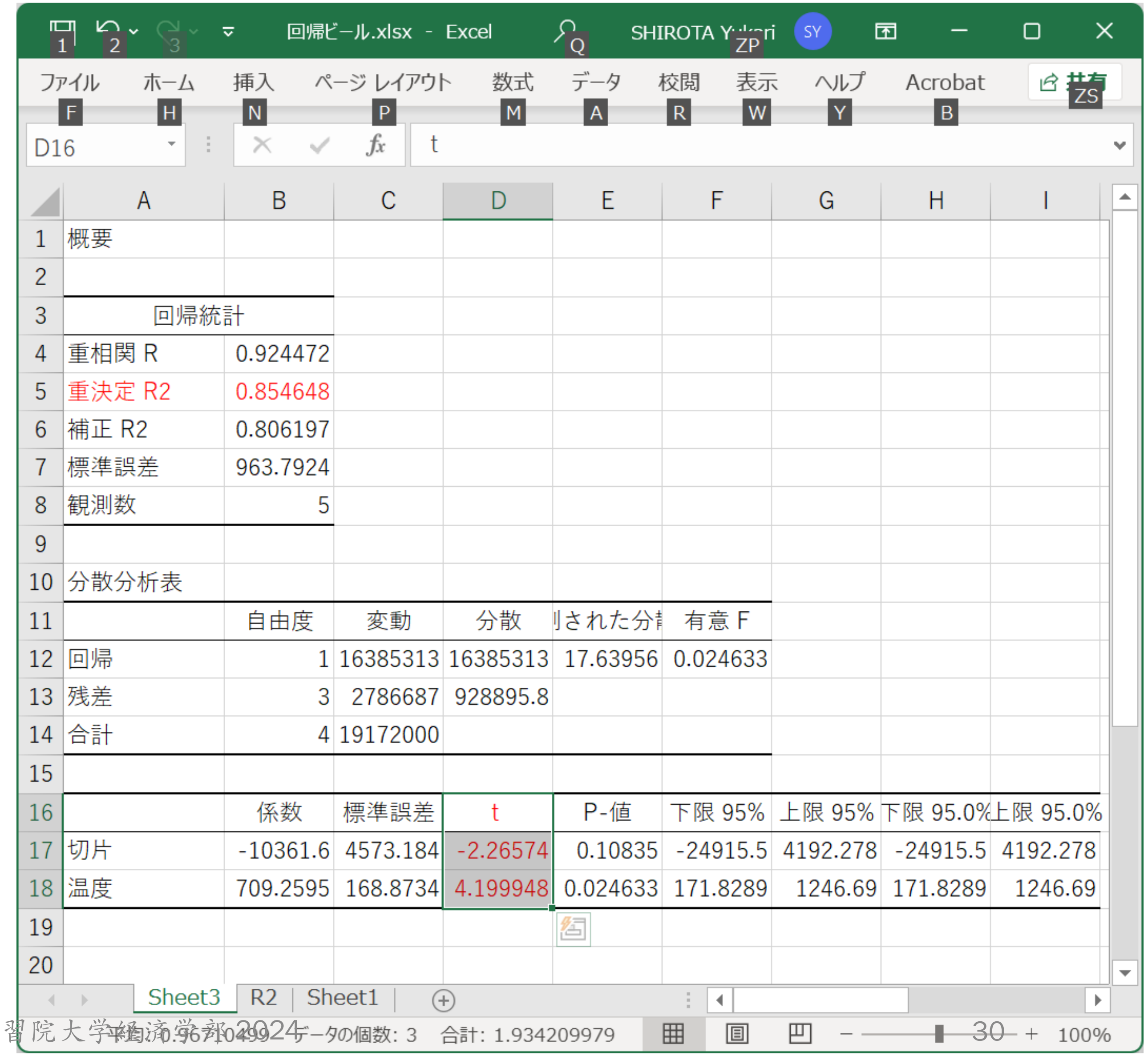
©白田由香利、学習院大学経済学部2024

回帰分析の結果の t値とは

- 考え方：説明変数xがyに全く関与していないならば，係数Bは０になるだろう．
- 帰無仮説　B=0
- 対立仮説　B≠0

$$t=\frac{b-0}{\sqrt{\dfrac{s^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}}$$

- この t 値は (n-2)の自由度のt-分布に従う．
- 4.1999は分布の裾野のほう



概要

回帰統計	
重相関 R	0.924472
重決定 R2	0.854648
補正 R2	0.806197
標準誤差	963.7924
観測数	5

分散分析表

	自由度	変動	分散	された分散	有意 F
回帰	1	16385313	16385313	17.63956	0.024633
残差	3	2786687	928895.8		
合計	4	19172000			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-10361.6	4573.184	-2.26574	0.10835	-24915.5	4192.278	-24915.5	4192.278
温度	709.2595	168.8734	4.199948	0.024633	171.8289	1246.69	171.8289	1246.69

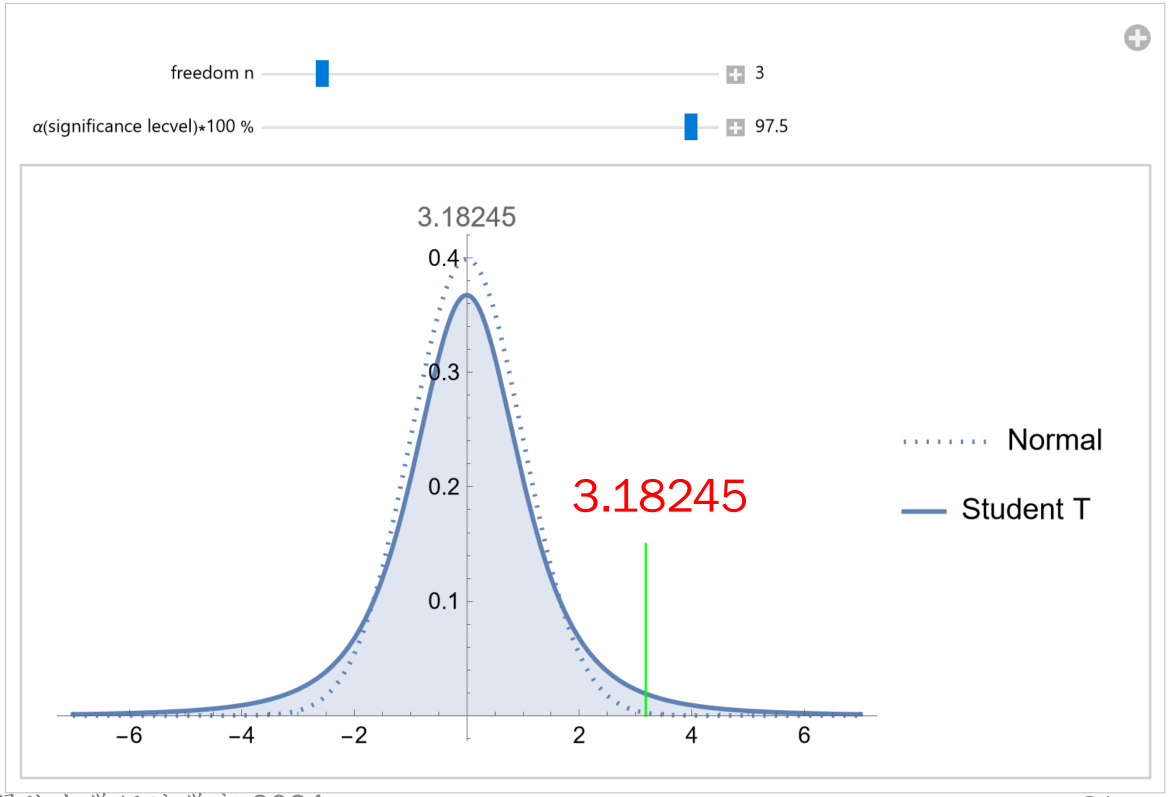
©白田由香利、学習院大学経済学部2024

回帰分析の結果の t値とは

- $t = \dfrac{b-0}{\sqrt{\dfrac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}$
- この t 値は
(5-2=3)の自由度のt-分布に従う.
5%の両側検定のとき,
境界値は3.18であることが
分布図から分かる.
- 今回の4.1999は帰無仮説棄却
領域に落ちる→帰無仮説棄却
- 説明変数 x は y に関係している
- 説明変数 x の効果はあった.

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-10361.6	4573.184	-2.26574	0.10835	-24915.5	4192.278	-24915.5	4192.278
温度	709.2595	168.8734	4.199948	0.024633	171.8289	1246.69	171.8289	1246.69

平均: 0.96710499　データの個数: 3　合計: 1.934209979



©白田由香利、学習院大学経済学部 2024

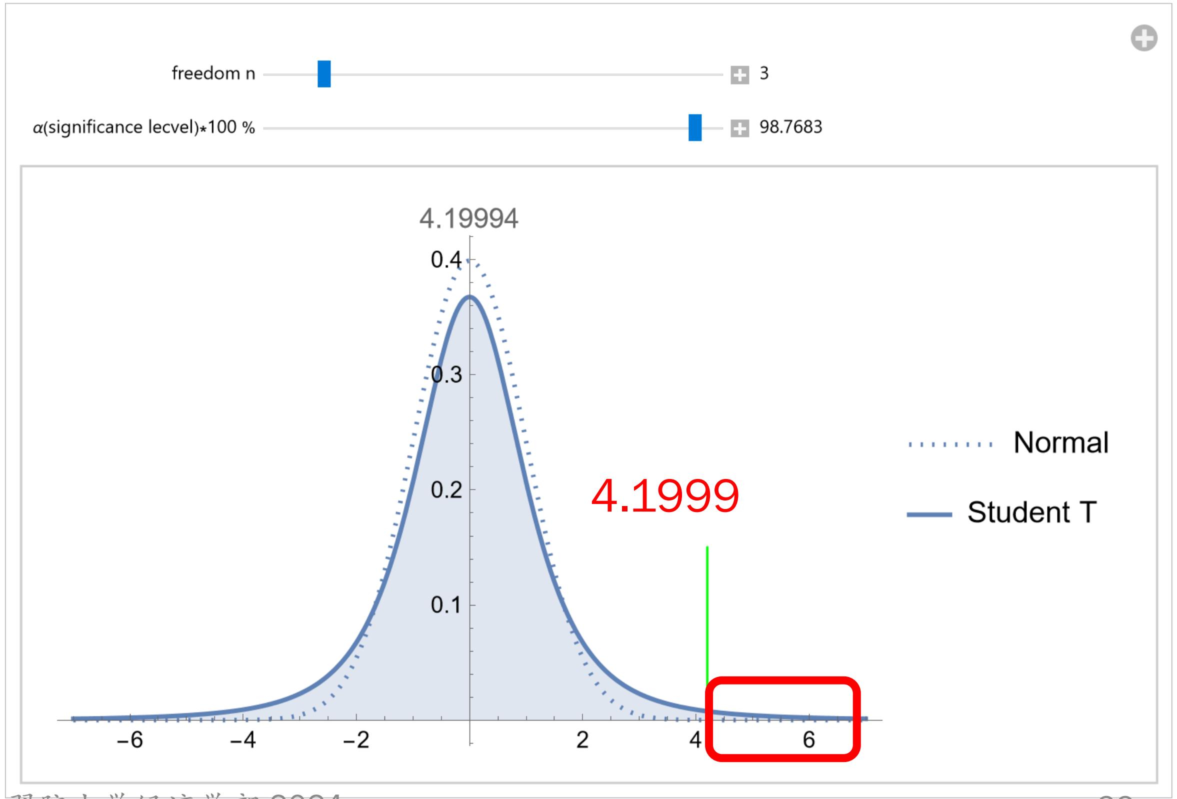
回帰分析の結果の t値とは

行		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	-10361.6	4573.184	-2.26574	0.10835	-24915.5	4192.278	-24915.5	4192.278
18	温度	709.2595	168.8734	4.199948	0.024633	171.8289	1246.69	171.8289	1246.69
19									
20									

Sheet3 R2 Sheet1

平均: 0.96710499 データの個数: 3 合計: 1.934209979 100%

- **P－値** 0.0246＝2.46%
- 今回の t 値4.1999がどの位ありえない値であるかを示す
- 両側検定なので，2で割り 1.23%
- 4.1999以上の値となる確率は1.23%である．面積小さい
- 有意水準を５%にとれば，５＞2.46 であるので，棄却領域におちることが分かる



©白田由香利、学習院大学経済学部 2024

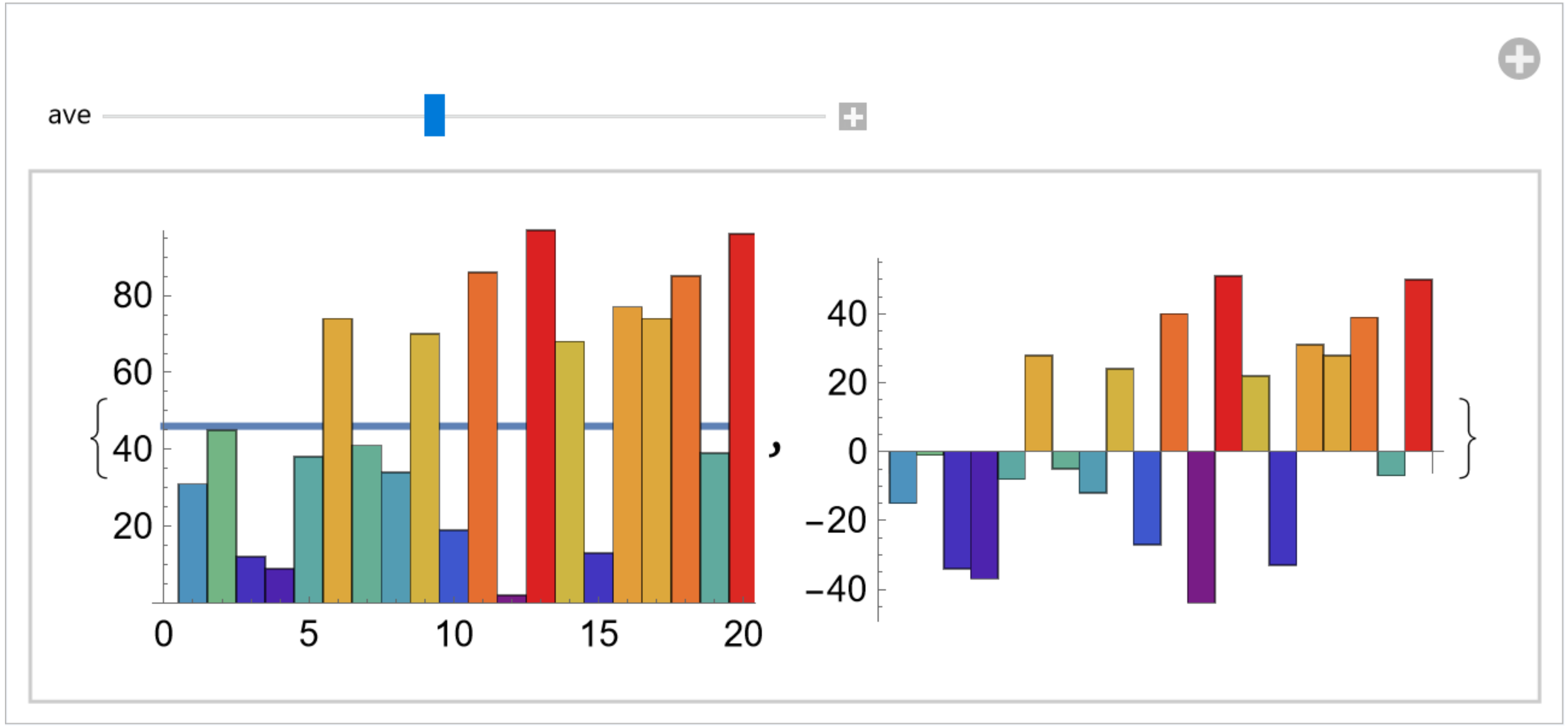
相関係数

A社の株価があがるとB社の株価が上がる場合，両者の間の相関は高い

©白田由香利、学習院大学経済学部 2024

平均 $\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- 同じ業種の20社の売上高
- 業界平均より上から下かが重大問題

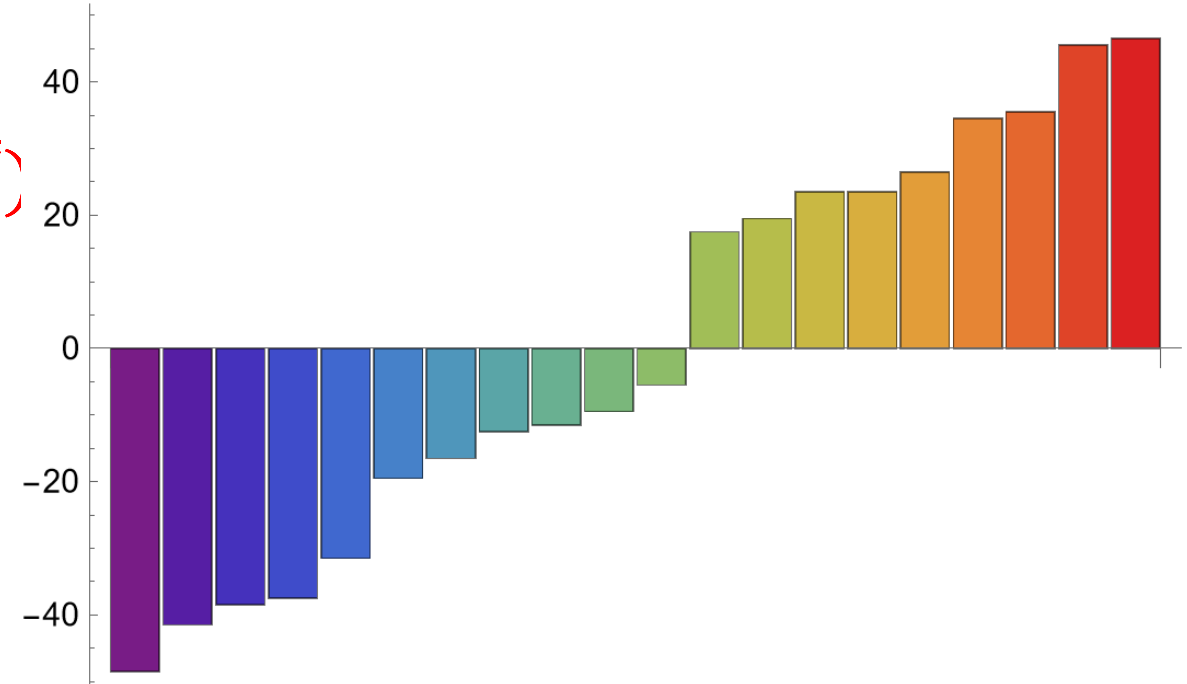


[グラフィクス教材](www-cc.gakushuin.ac.jp/~20010570/VDStat/)
www-cc.gakushuin.ac.jp/~20010570/VDStat/

©白田由香利、学習院大学経済学部 2024

偏差を合計したら０

- *平均の定義式を変形していくと*
- $\bar{X} = \frac{x_1+x_2+\cdots+x_n}{n}$
- *n=3でやってみる*
- $\bar{X} + \bar{X} + \bar{X} = x_1 + x_2 + x_3$
- $(x_1 - \bar{X}) + (x_2 - \bar{X}) + (x_3 - \bar{X})$
- $\sum(\mathrm{x}_i - \bar{X}\ ) = 0$



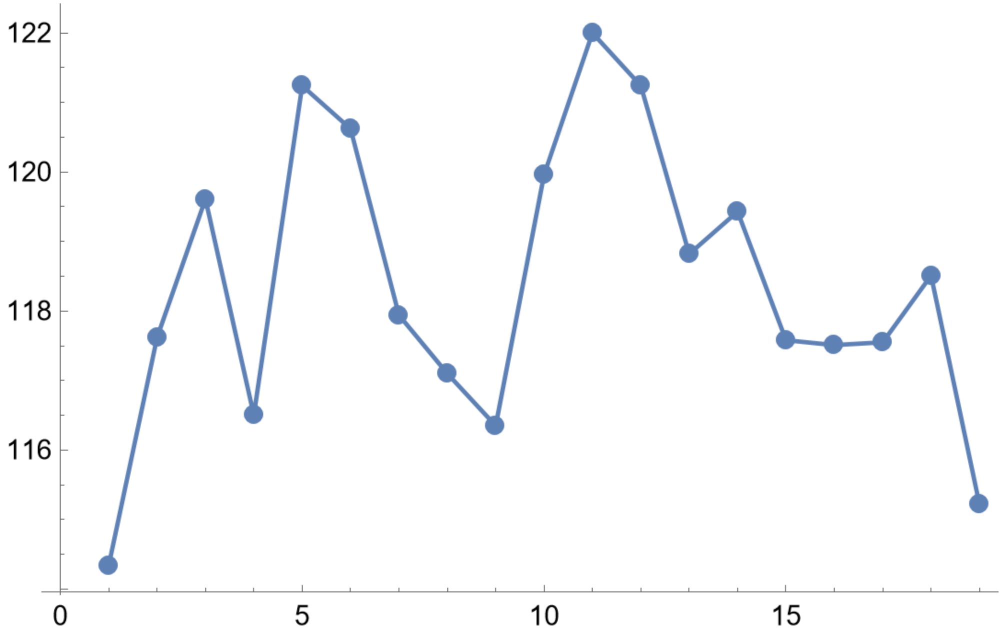
分散：偏差の平方和を自由度で割った値

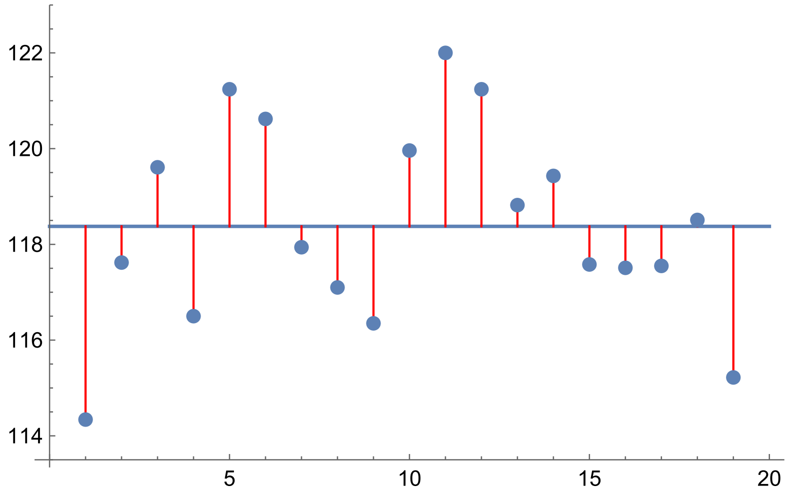
全データが自由に動ける場合（記述統計）

$$Var(X) = \frac{1}{n}\sum(x_i - \bar{X})^2$$

- 揺れ幅の大きさ
- 株価の場合，リスクが大きい（上下動が大きい）

- 平均を求めて
- 偏差を求める
- それを平方して合計して割る

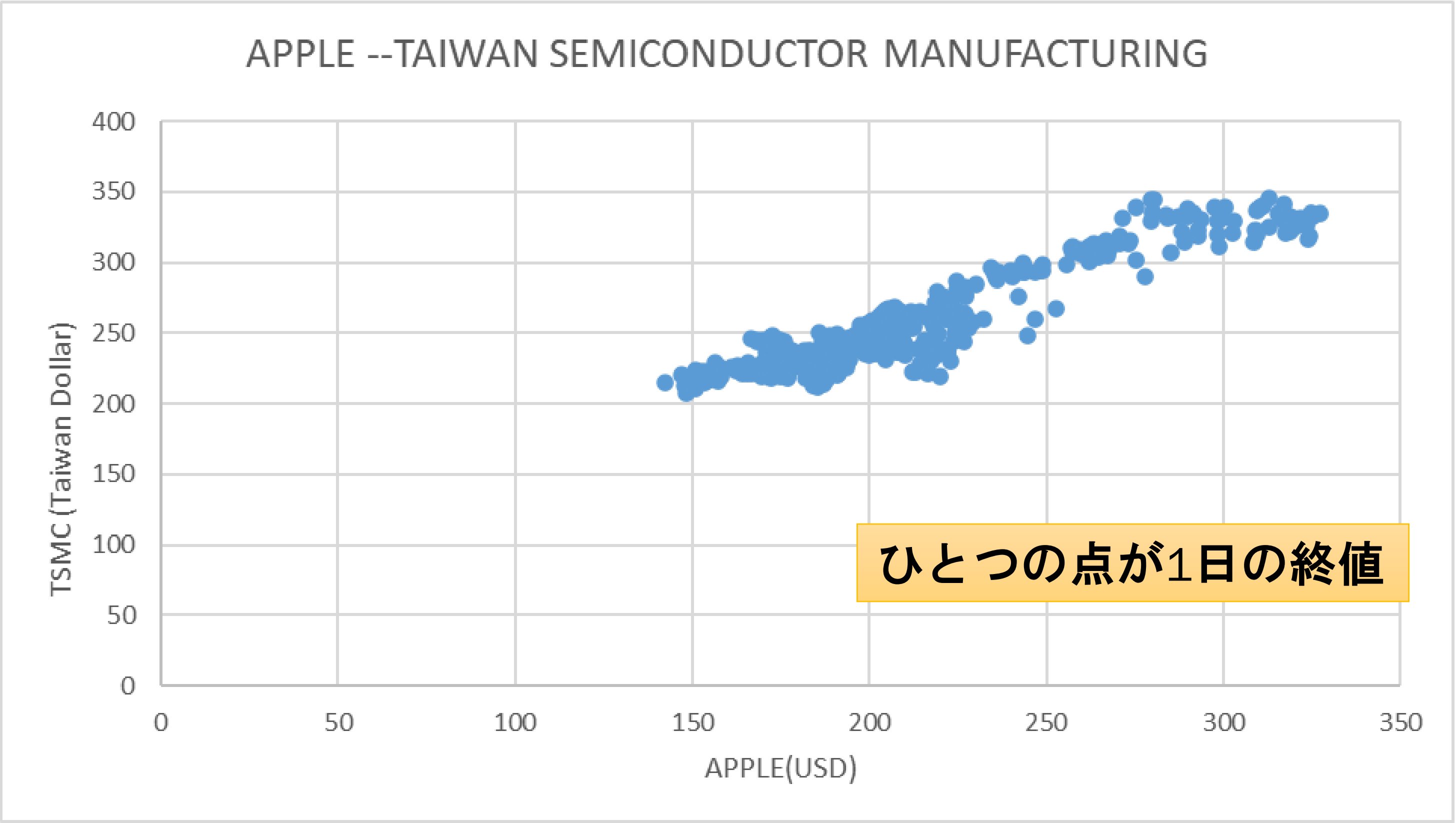




2018/4/2 – 2020/3/19 相関係数

	APPLE INC.	SAMSUNG ELECTRONICS CO.,LTD.	HON HAI PRECISION INDUSTRY CO., LTD.	HITACHI, LTD.	SONY CORPORATION	PANASONIC CORPORATION	INTEL CORP	HP INC.	LG ELECTRONICS INC.	CISCO SYSTEMS INC	PEGATRON CORPORATION	MITSUBISHI ELECTRIC CORPORATION	GENERAL DYNAMICS CORP	FUJITSU LIMITED	TAIWAN SEMICONDUCTOR MANUFACTURING COMPANY LIMITED
APPLE INC.	1.00														
SAMSUNG ELECT	0.81	1.00													
HON HAI PRECISI	0.10	0.37	1.00												
HITACHI, LTD.	0.58	0.75	0.48	1.00											
SONY CORPORAT	0.90	0.68	0.09	0.54	1.00										
PANASONIC CORP	-0.16	0.17	0.83	0.18	-0.10	1.00									
INTEL CORP	0.69	0.84	0.26	0.62	0.52	0.10	1.00								
HP INC.	-0.10	-0.04	0.45	-0.07	0.05	0.69	-0.09	1.00							
LG ELECTRONICS	-0.34	0.13	0.68	0.32	-0.39	0.70	0.09	0.24	1.00						
CISCO SYSTEMS I	-0.07	-0.16	-0.35	0.12	-0.15	-0.48	-0.04	-0.20	-0.05	1.00					
PEGATRON CORP	0.46	0.68	0.86	0.66	0.43	0.66	0.56	0.30	0.51	-0.39	1.00				
MITSUBISHI ELEC	0.27	0.59	0.70	0.67	0.22	0.59	0.52	0.26	0.67	0.02	0.74	1.00			
GENERAL DYNAM	-0.01	0.25	0.74	0.47	0.09	0.72	0.10	0.49	0.64	-0.12	0.62	0.70	1.00		
FUJITSU LIMITED	0.93	0.77	-0.09	0.48	0.82	-0.32	0.71	-0.23	-0.43	-0.01	0.31	0.14	-0.17	1.00	
TAIWAN SEMICON	0.94	0.80	0.05	0.61	0.84	-0.28	0.71	-0.28	-0.30	0.01	0.44	0.28	-0.06	0.93	1.00

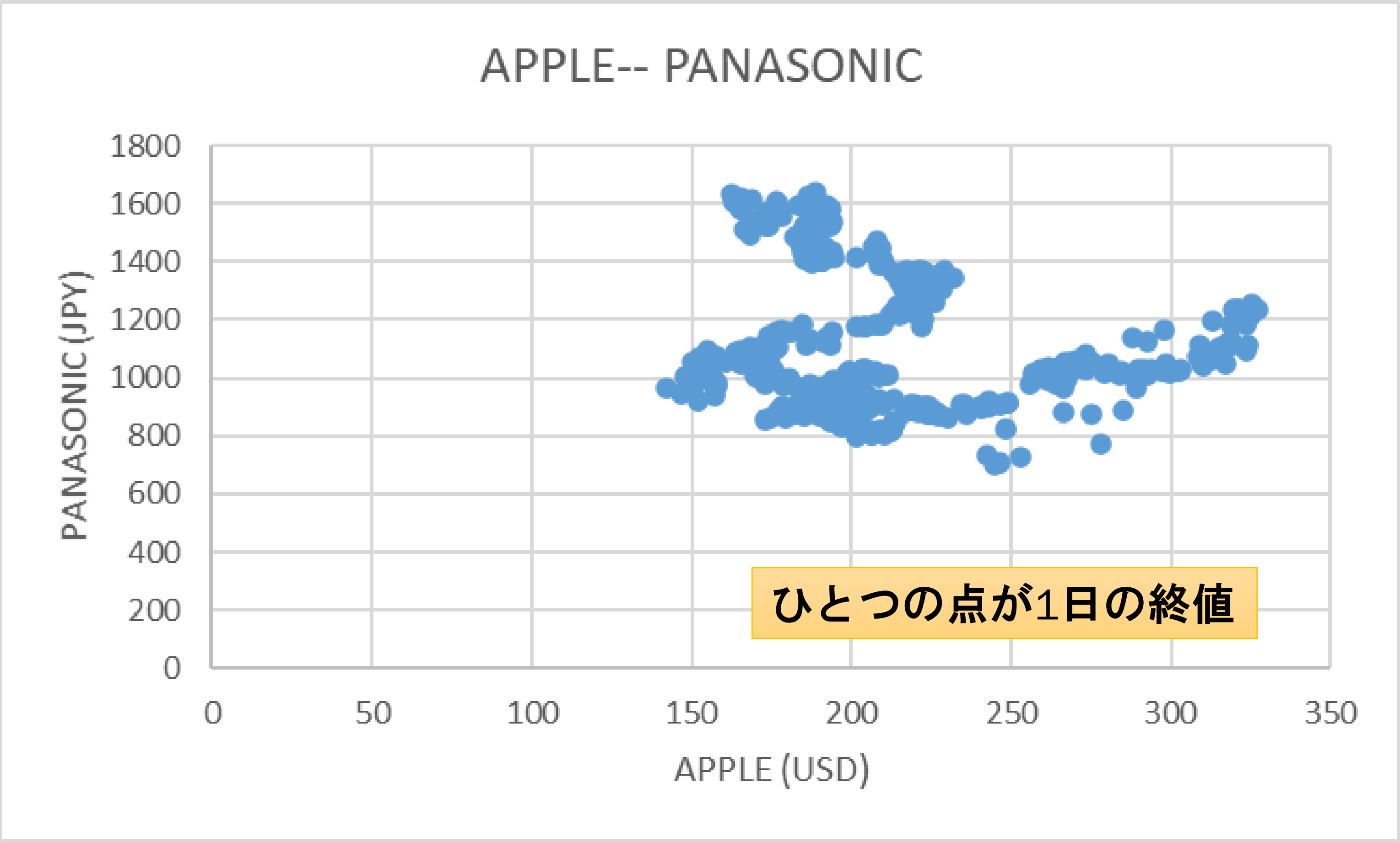
APPLEとTSMCの株価の相関
散布図でひとめで相関が高いと分かる。0.94



A社が上がると，B社もあがる
A社が下がると，B社も下がる

どちらが主導権をとっているかは
この図と相関係数からでは不明
両社の間には取引は全くないかも
しれない

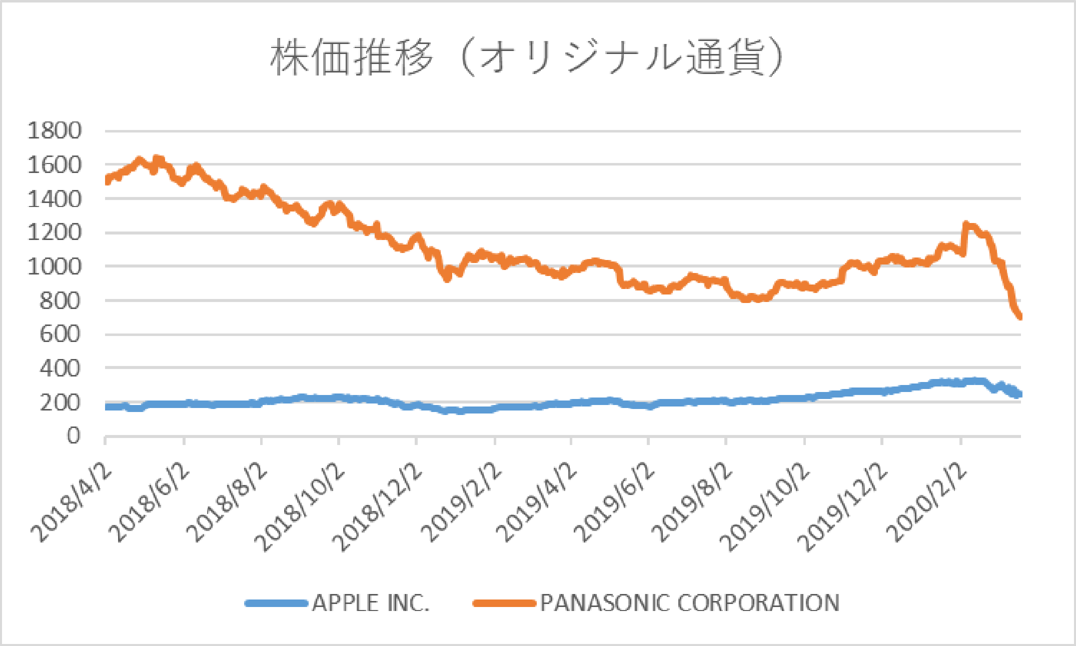
APPLEとPANASONIC
相関係数は－0.16で若干負となった。

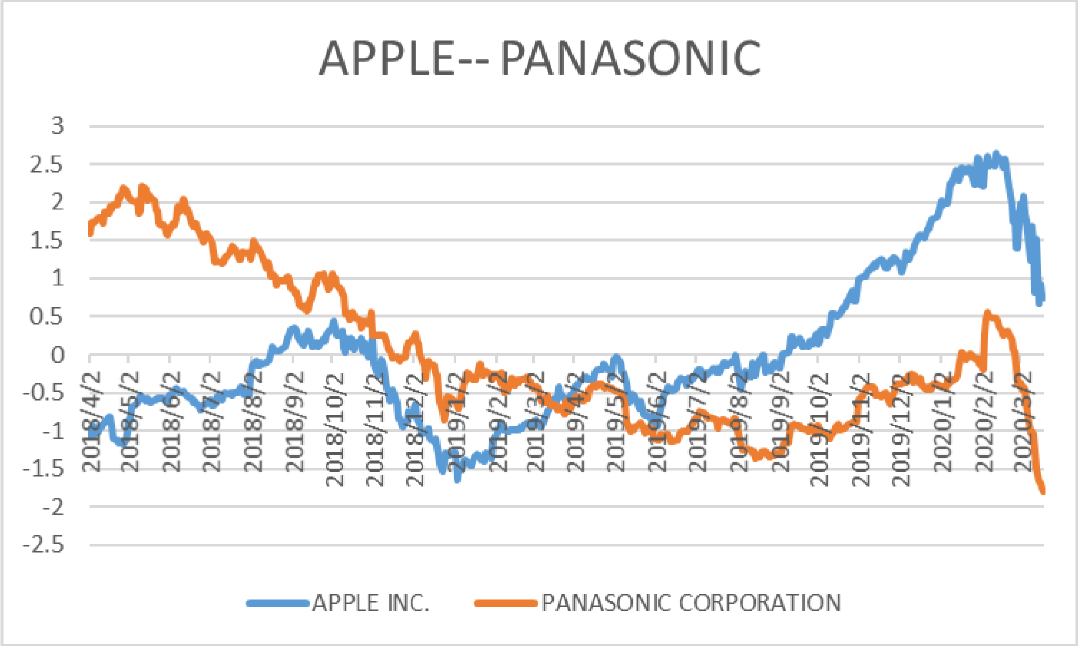


株価の変動パターン比較は標準化して行う

標準化：対象期間の中で，平均が０，分散が１となるようにする

平均を計算して，各データから引き算する．→偏差
分散の平方根（標準偏差）を計算して，各データの偏差を標準偏差で割る

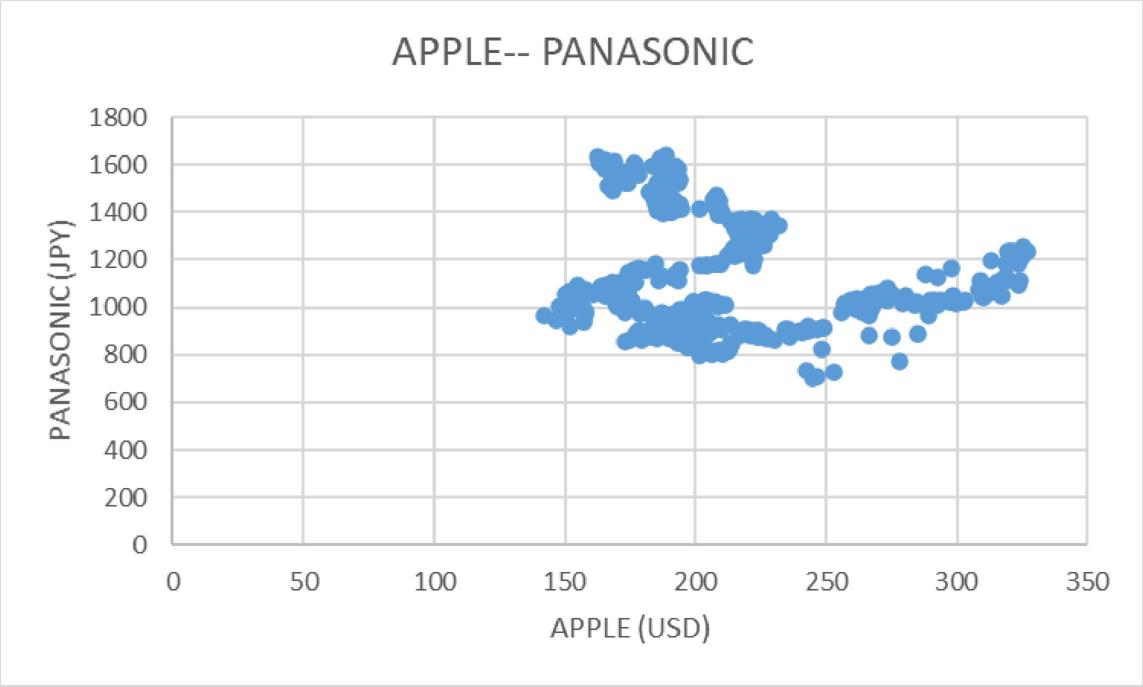


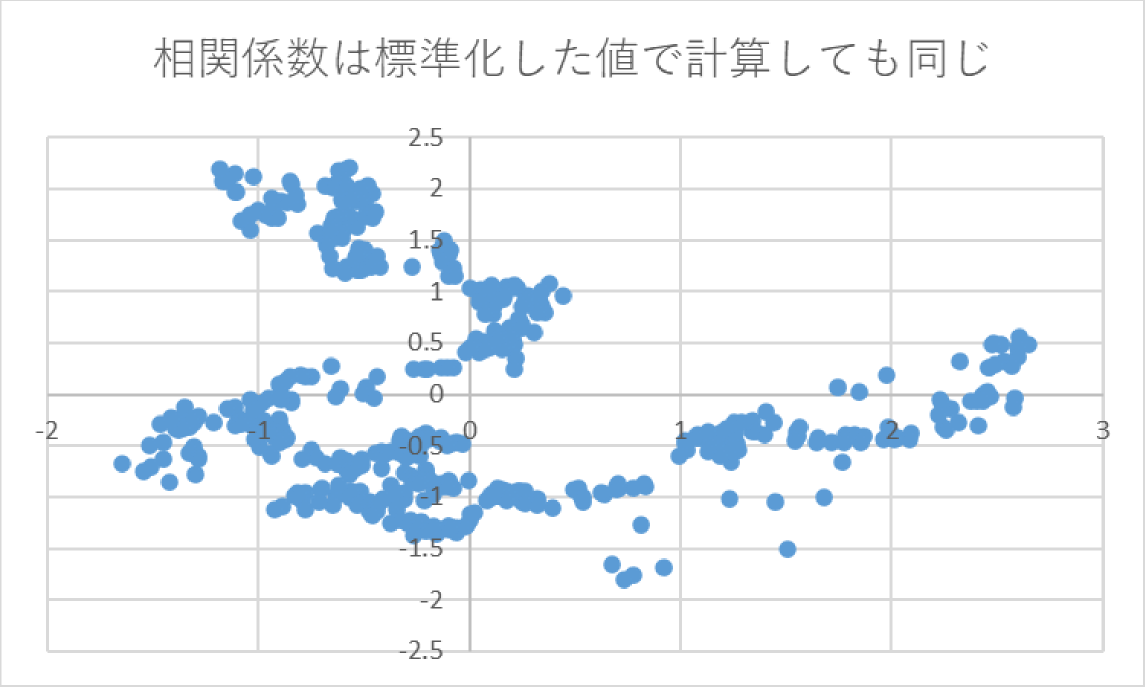


2年間、APPLEは伸び、パナは下降している様子が見える

株価の変動パターン比較は標準化して行う

- 標準化した値でも相関係数は同じになる





標準化：xi → zi 平均を０，分散を１に

統計の公式

分散＝（データの2乗の平均値）－（平均値の2乗）

証明

$$S^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 = \frac{1}{n}\Sigma(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \frac{1}{n}\Sigma x_i^2 - \frac{1}{n}\cdot 2\cdot\bar{x}\cdot\Sigma x_i + \frac{1}{n}\underbrace{\Sigma\bar{x}^2}_{n\times\bar{x}^2}$$

$$= \frac{1}{n}\Sigma x_i^2 - 2\cdot\bar{x}\cdot\bar{x} + \bar{x}^2$$

$$= \frac{1}{n}\Sigma x_i^2 - (\bar{x})^2$$

$$= \overline{x^2} - (\bar{x})^2$$

標準化：xi → zi 平均を０，分散を１に

標準化　$x_i \to z_i = \dfrac{x_i - \bar{x}}{S_x}$ ← xの標準偏差.

$$\bar{z} = \frac{1}{n}\sum z_i = \frac{1}{n}\cdot\frac{1}{S_x}\underbrace{\left(\sum x_i - \sum \bar{x}\right)}_{\text{偏差の和なので}0} = 0.$$

$$Var(z) = \overline{z^2} - \underbrace{(\bar{z})^2}_{\text{平均}0\text{なので}0}$$

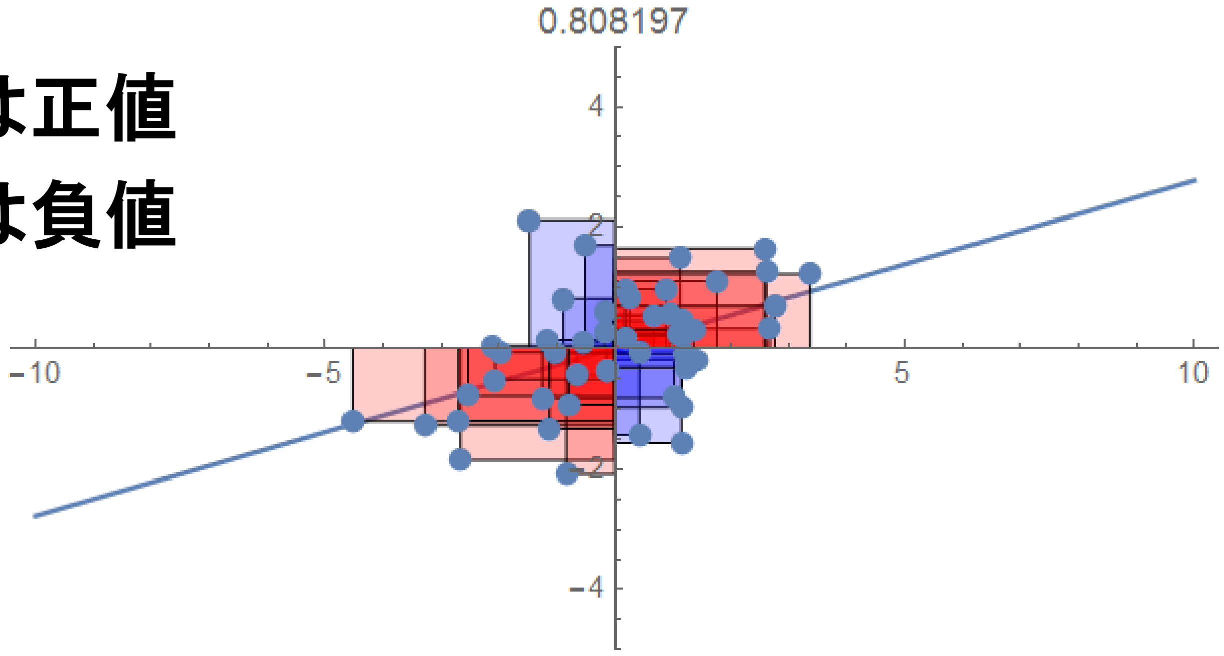
$$= \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{S_x}\right)^2 = \frac{1}{S_{xx}} \times \underbrace{\frac{1}{n}\sum(x_i - \bar{x})^2}_{S_{xx}}$$

$$= \frac{S_{xx}}{S_{xx}} = 1$$

相関係数は共分散から求める
共分散の可視化による説明

- 偏差の正負によってその項の正負が違う
- 両方プラスOR両方マイナス⇒　赤
- プラスとマイナス⇒青
- 赤のパネルが広ければ、共分散は正値
- 青のパネルが広ければ、共分散は負値

赤パネルの面積合計のほうが広いので，共分散は正の値　0.808



$$S_{xy} = \frac{\sum(x_i - \bar{X}) \times (y_i - \bar{Y})}{n}$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

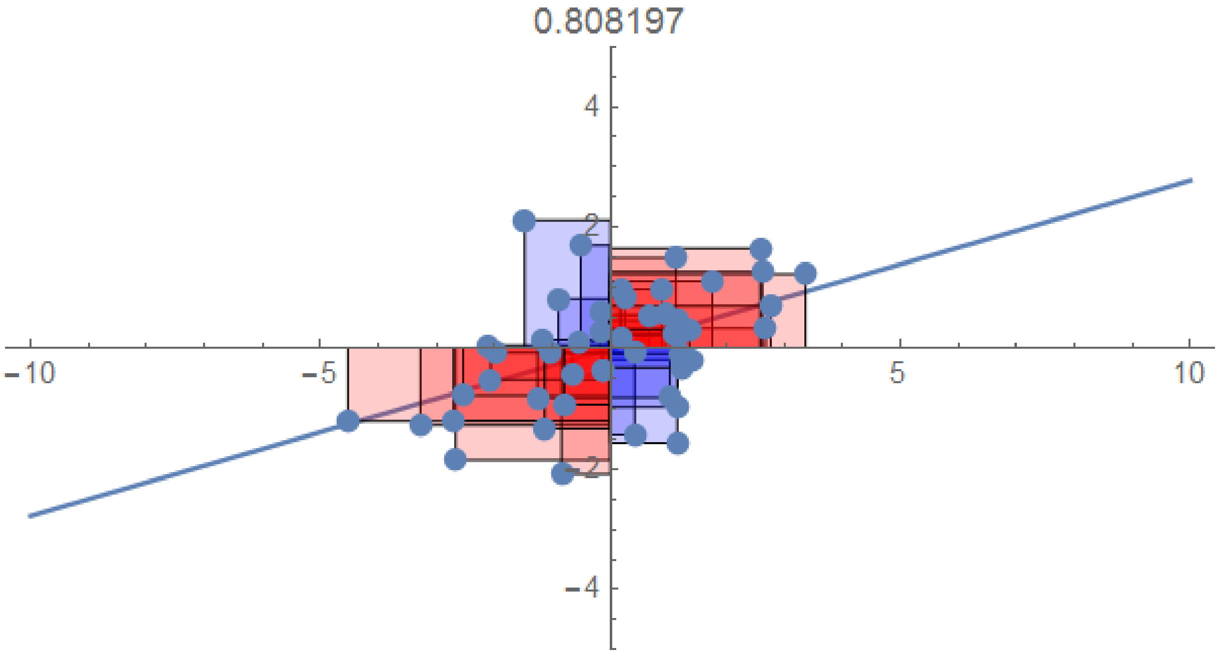
相関係数
－１から１までの値をとる

- －１の場合，正反対の動き
- ＋１の場合，全く同じ動き

$$S_{xy} = \frac{\sum(x_i - \bar{X}) \times (y_i - \bar{Y})}{n}$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$\frac{\frac{\sum(x_i-\bar{X})\times(y_i-\bar{Y})}{n}}{\sqrt{S_{xx}S_{yy}}} = \frac{\frac{\sum(x_i-\bar{X})\times(y_i-\bar{Y})}{n}}{\sqrt{\frac{\sum x_i^2}{n}\times\frac{\sum y_i^2}{n}}}$$



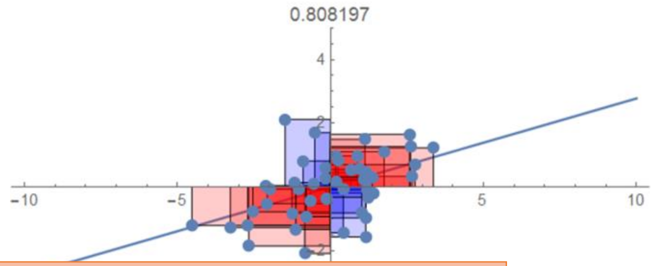
©白田由香利、学習院大学経済学部 2024

公式
線形回帰の傾きは共分散から求められる

$$\begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix}$$ 共分散行列と呼ぶ

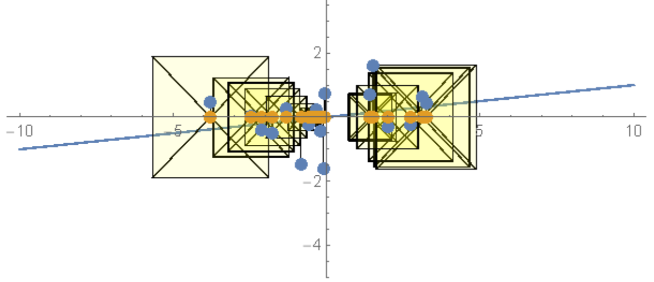
$$\left(\begin{array}{c} \text{右図は予め重心を原点に移動しておく} \\ set \quad \bar{x} = 0, \quad \bar{y} = 0 \end{array} \right)$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\frac{\sum x_i \times y_i}{n}}{\frac{\sum (x_i)^2}{n}}$$



分子:共分散値

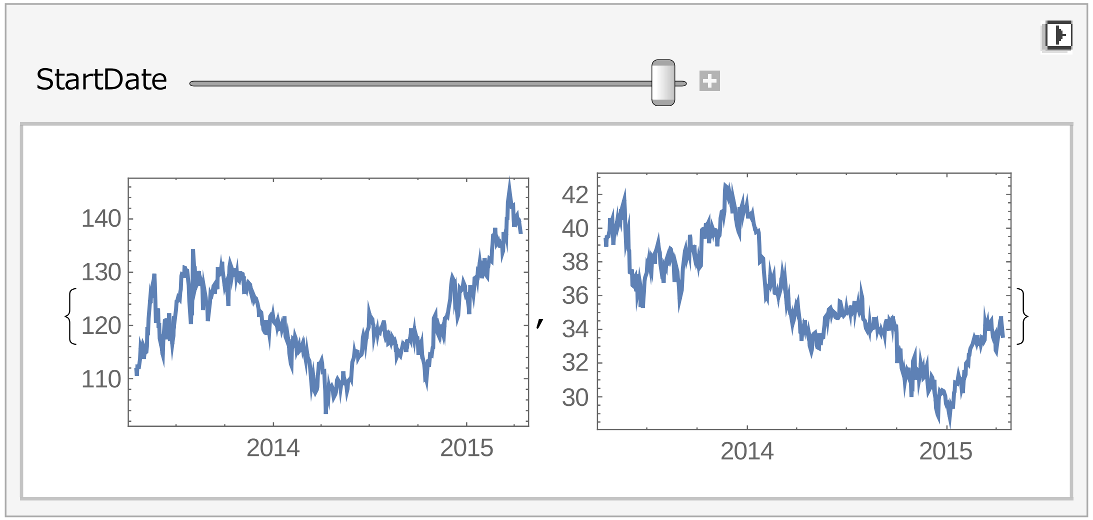
分母：ｘの分散値



共分散の項を、ｘの分散で割ることで、スケーリングしている。
Xの変化に対して、どの程度の影響があるかを見たいため。

共分散は2社の株価時系列データ$\{x_i\}$ と $\{y_i\}$ の類似度を表す（i日目の株価）

Stock Price Time Series Data $\{x_i\}$ and $\{y_i\}$



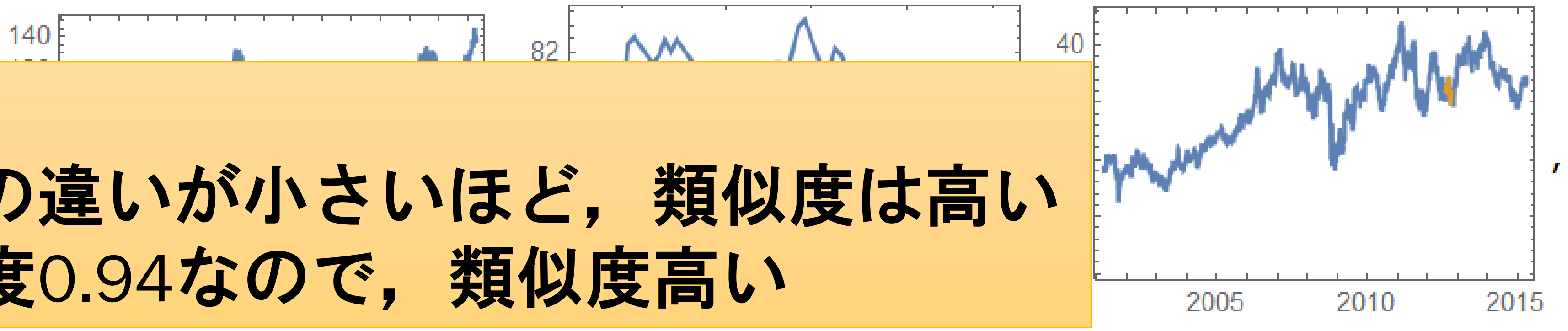
始めに標準化しておく

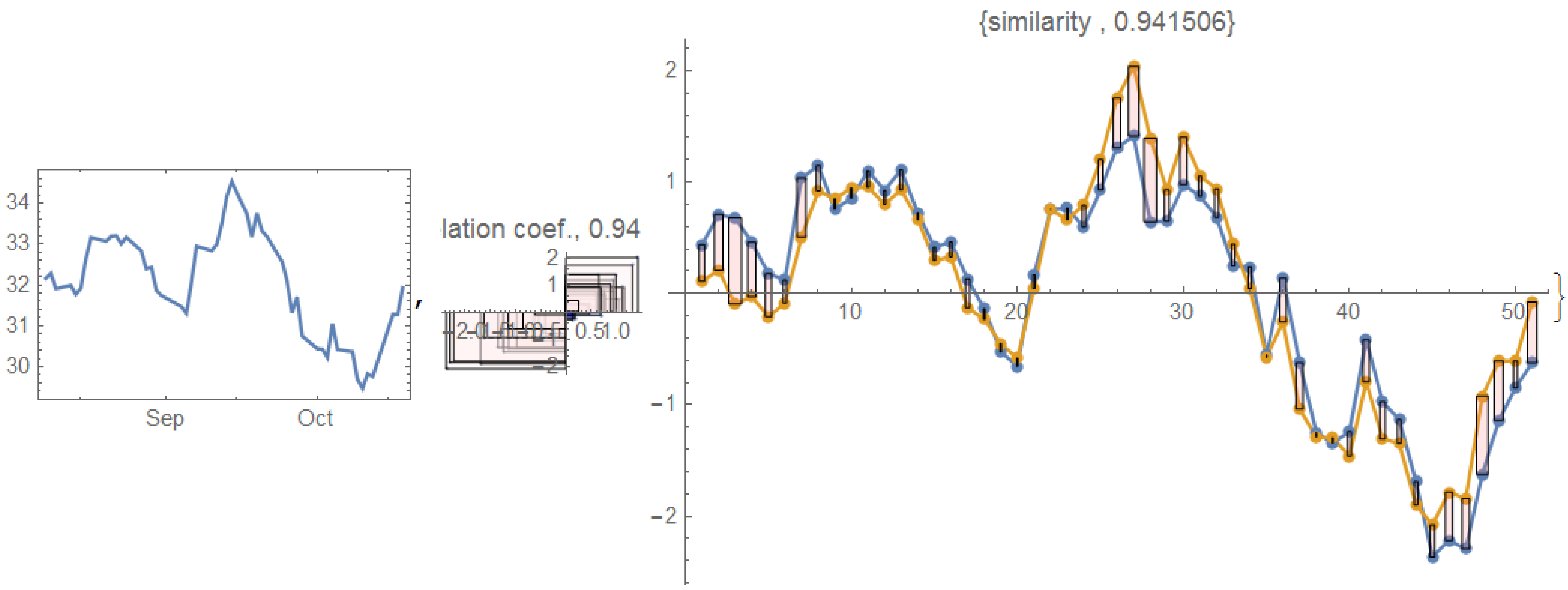
$$\bar{x} = 0, \quad \bar{y} = 0, \quad \sqrt{\frac{\sum(x_i)^2}{n}} = 1, \quad \sqrt{\frac{\sum(y_i)^2}{n}} = 1$$

Reference: Private communication with Prof Tetsuji Kuboyama, Gakushuin Univ.

©白田由香利、学習院大学経済学部 2024

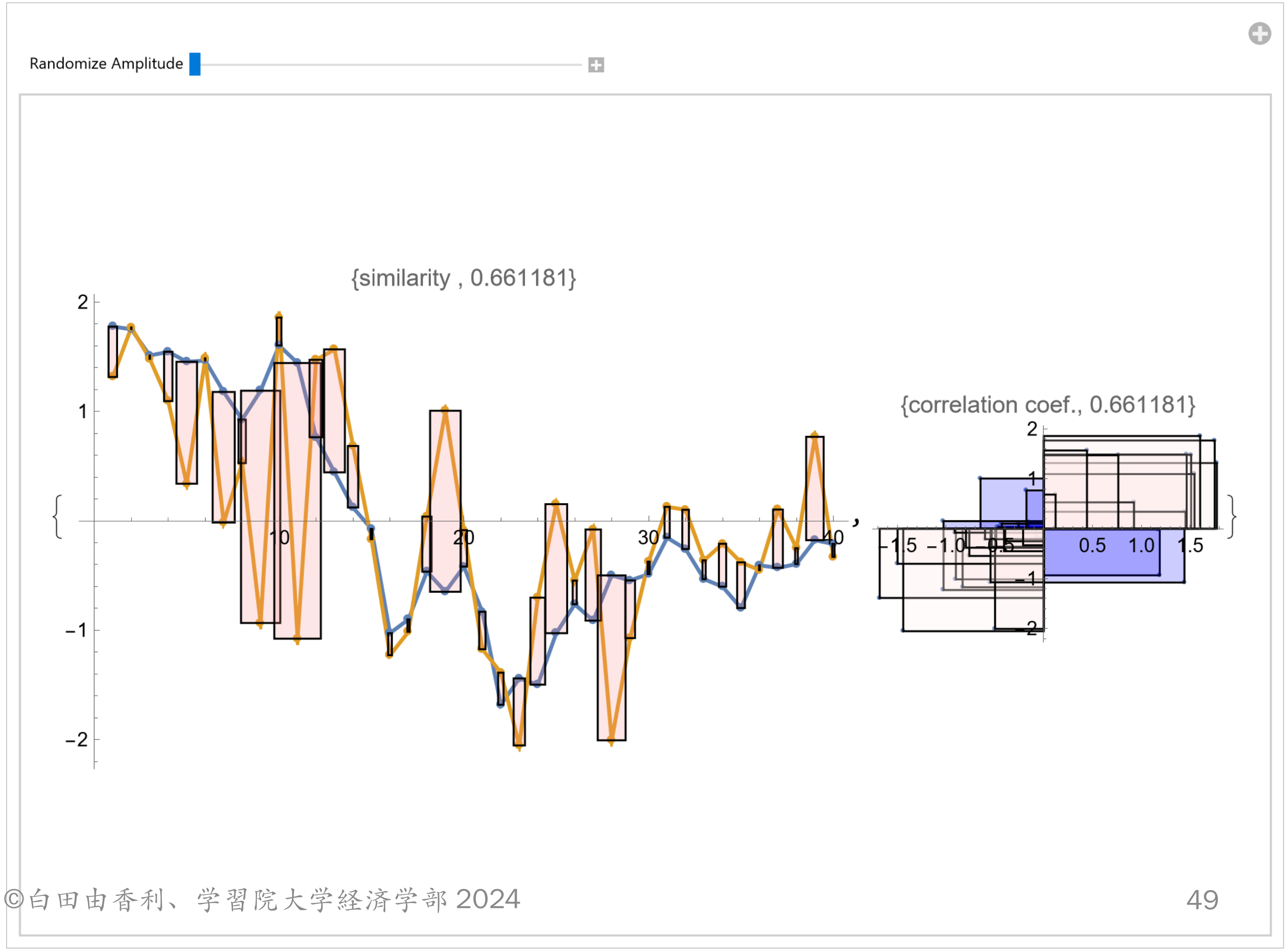
株価の違いが小さいほど，類似度は高い
類似度0.94なので，類似度高い





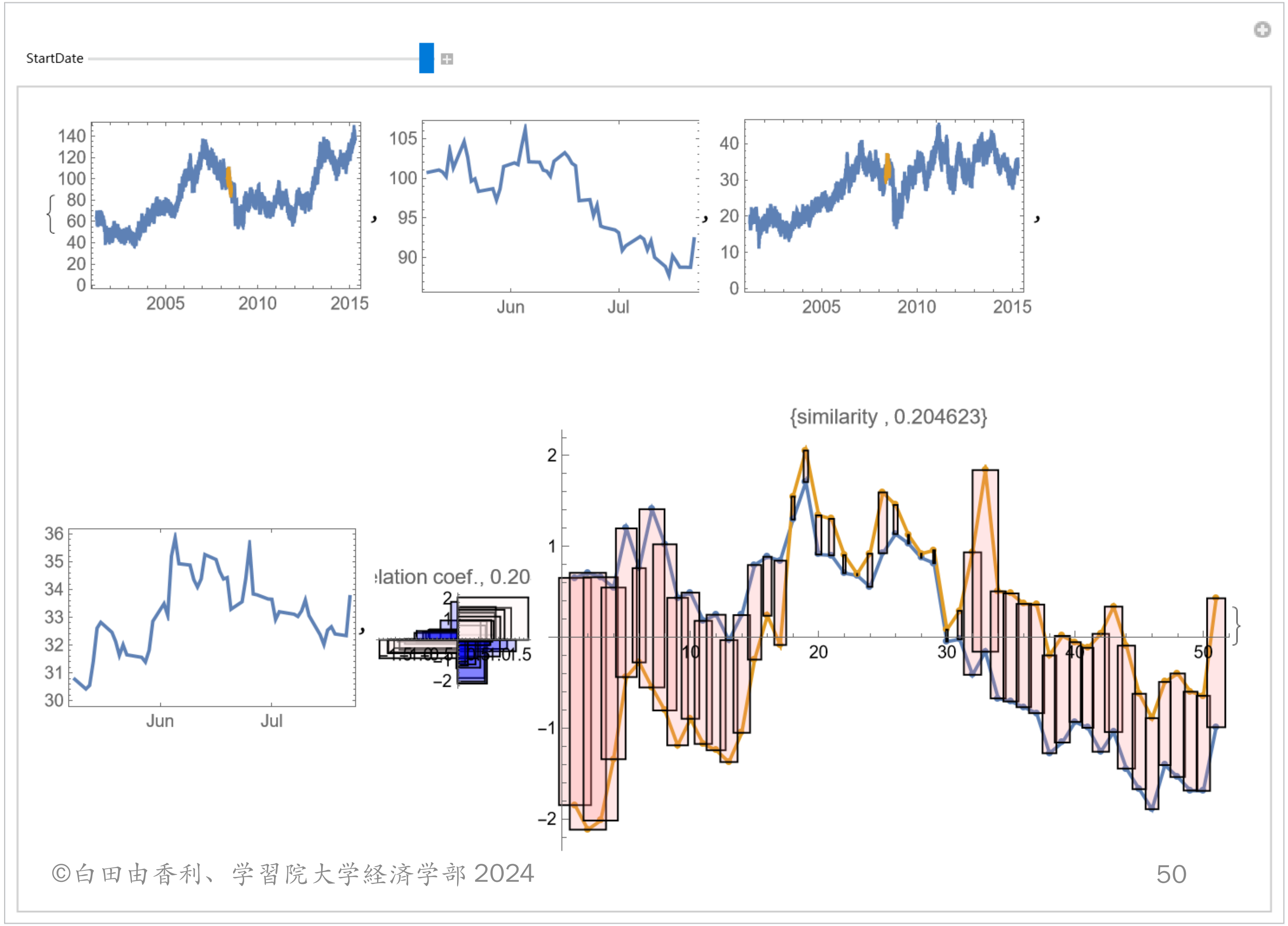
グラフィクス教材
www-cc.gakushuin.ac.jp/~20010570/VDStat/

- 標準化してあるので共分散＝相関係数
- 相関係数が大きいと類似度は大きい
- 右の例では0.66ずれが目立つ



グラフィクス教材
www-cc.gakushuin.ac.jp/~20010570/VDStat/

- X社とY社の株価変動の類似度
- どの期間をとるかで類似度は変わる



Similarity: Let's calculate the sum of squared distances

$$\sum (x_i - y_i)^2$$

Larger the sum of the squared distances, then smaller the similarity becomes.

$$= \sum (x_i{}^2 - 2x_i y_i + y_i{}^2) = \sum x_i{}^2 - 2\sum x_i y_i + \sum y_i{}^2$$

$$= n - 2\sum x_i y_i + n = 2n - 2\sum x_i y_i = 2\left(n - \sum x_i y_i\right)$$

Then

$$\sum x_i y_i = n - \frac{1}{2}\sum (x_i - y_i)^2$$

The Correlation Coefficient which shows the similarity

$$\boxed{\frac{1}{n}\sum x_i y_i} = 1 - \frac{1}{2n}\underbrace{\sum (x_i - y_i)^2}_{\text{Sum of squared distance}}$$

if every $x_i = -y_i$ then $1 - \frac{1}{2n}\sum(-2y_i)^2 = 1 - \frac{1}{2n} \cdot 4n = \underline{-1}$

©白田由香利、学習院大学経済学部 2024

キーワード

- 平均　$\bar{X} = \frac{x_1+x_2+\cdots+x_n}{n}$
- 分散：偏差の平方和を自由度で割った値
  - 全データが自由に動ける場合（記述統計）$Var(X) = \frac{1}{n}\sum(x_i - \bar{X})^2$
  - 推測統計の場合、母平均の代わりに標本平均を用いているので自由度が1減る。$Var(X) = \frac{1}{n-1}\sum(x_i - \bar{X})^2$

  - 標準偏差は、分散のルートを取った値（記述統計版と推測統計版）
- 共分散：変数x, yの関係を表す

$$S_{xy} = \frac{\sum(x_i-\bar{X})\times(y_i-\bar{Y})}{n-1}$$ (推測統計の場合）

- 相関係数　$\mathrm{r_{xy}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

　共分散を、xの標準偏差とｙ標準偏差で割った値

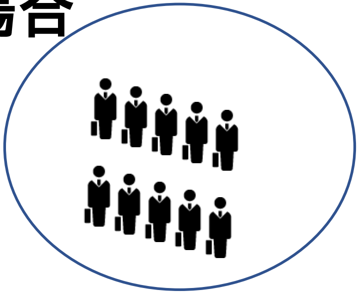
推測統計では分散の式は(n-1)で割る

- 分散の定義式
- ｎで割るのが記述統計→　母分散
- 推測統計では（n – 1)で割る→　不偏分散
- 分散の定義：偏差の平方和を自由度で割った値
- 記述統計と推測統計では自由度が異なる

記述統計VS 推測統計

記述統計

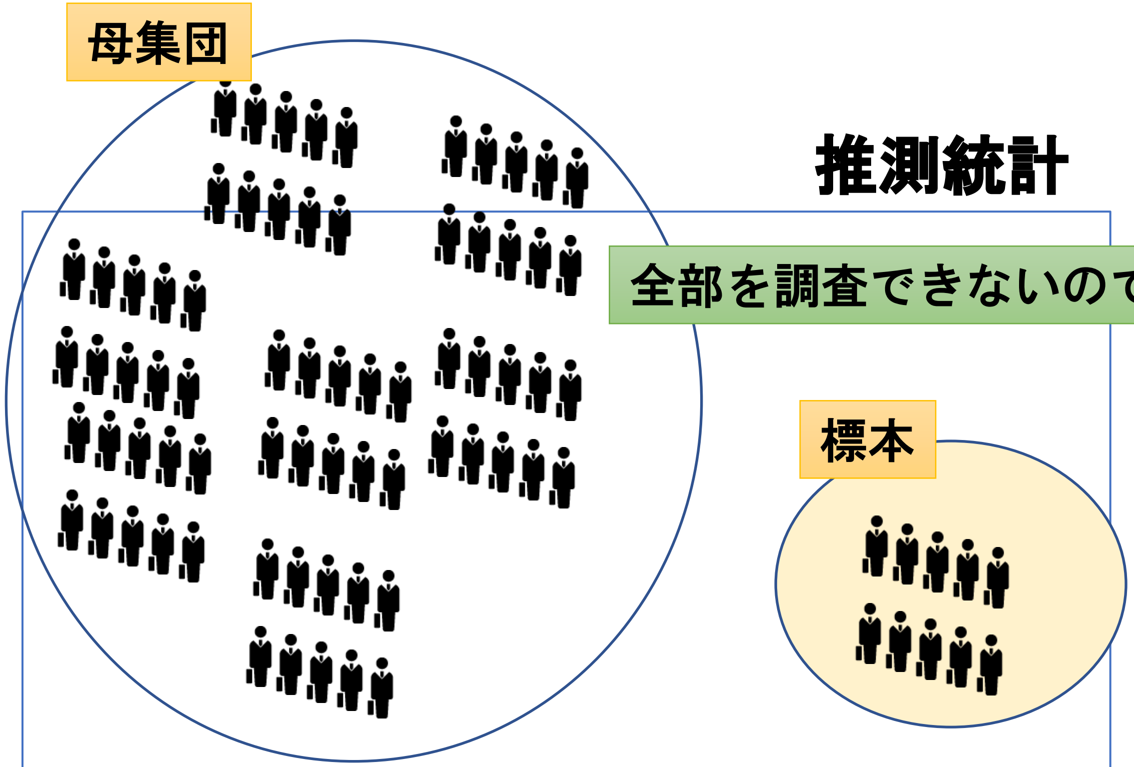
- これが母集団の場合



$$\mu = \frac{1}{10}\sum x_i \quad Var(X) = \frac{1}{10}\sum (x_i - \mu)^2$$

μは母平均

推測統計



全部を調査できないので標本

$$\bar{X} = \frac{1}{10}\sum x_i \quad Var(X) = \frac{1}{10-1}\sum (x_i - \bar{X})^2$$

標本平均と標本分散(不偏分散）
自由度は９

©白田由香利、学習院大学経済学部 2024

標本平均が制約となって自由度が減る

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- 標本が以下の値だったとする

標本平均の定義式を変形していくと

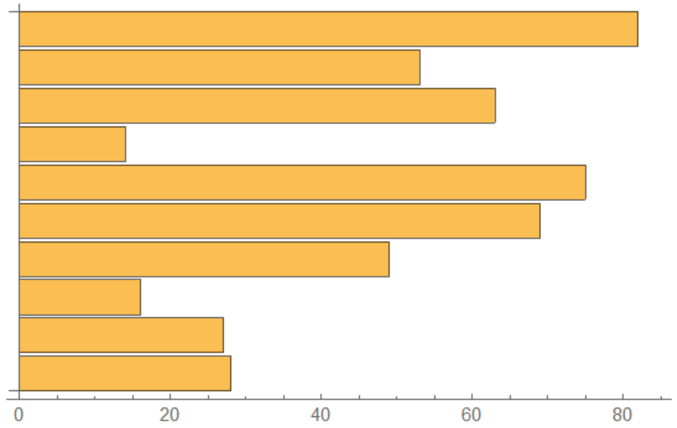
$\bar{X} + \bar{X} + \bar{X} = x_1 + x_2 + x_3$

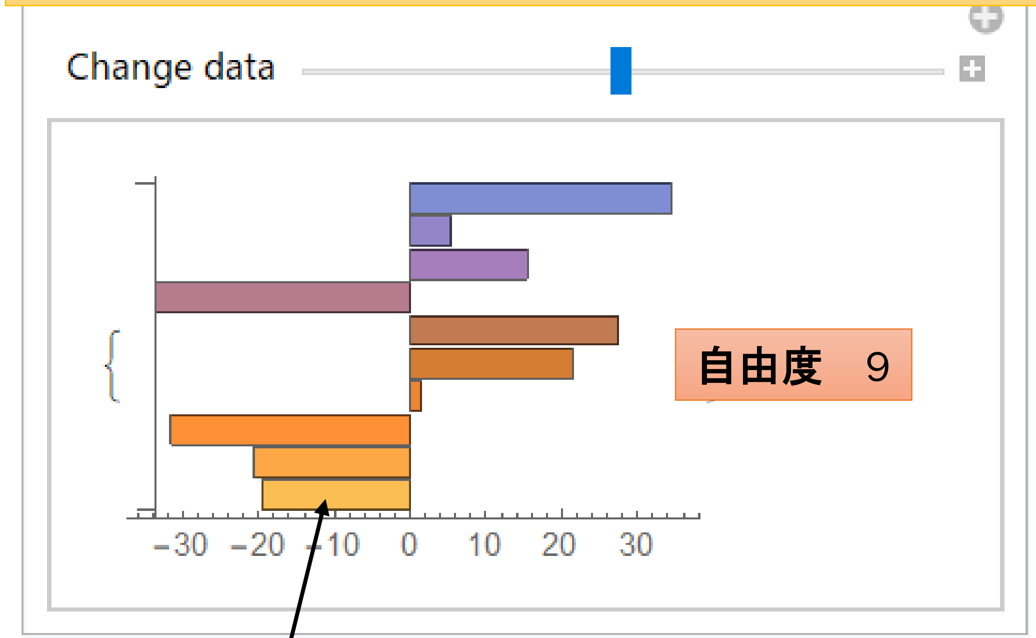
$(x_1 - \bar{X}) + (x_2 - \bar{X}) + (x_3 - \bar{X}) = 0$

$\sum(x_i - \bar{X}) = 0$

偏差の合計は０であるという制約が課せられている

よって、最後の1個は自由に動けない





自由度　9

最後の1個は自由に動けない

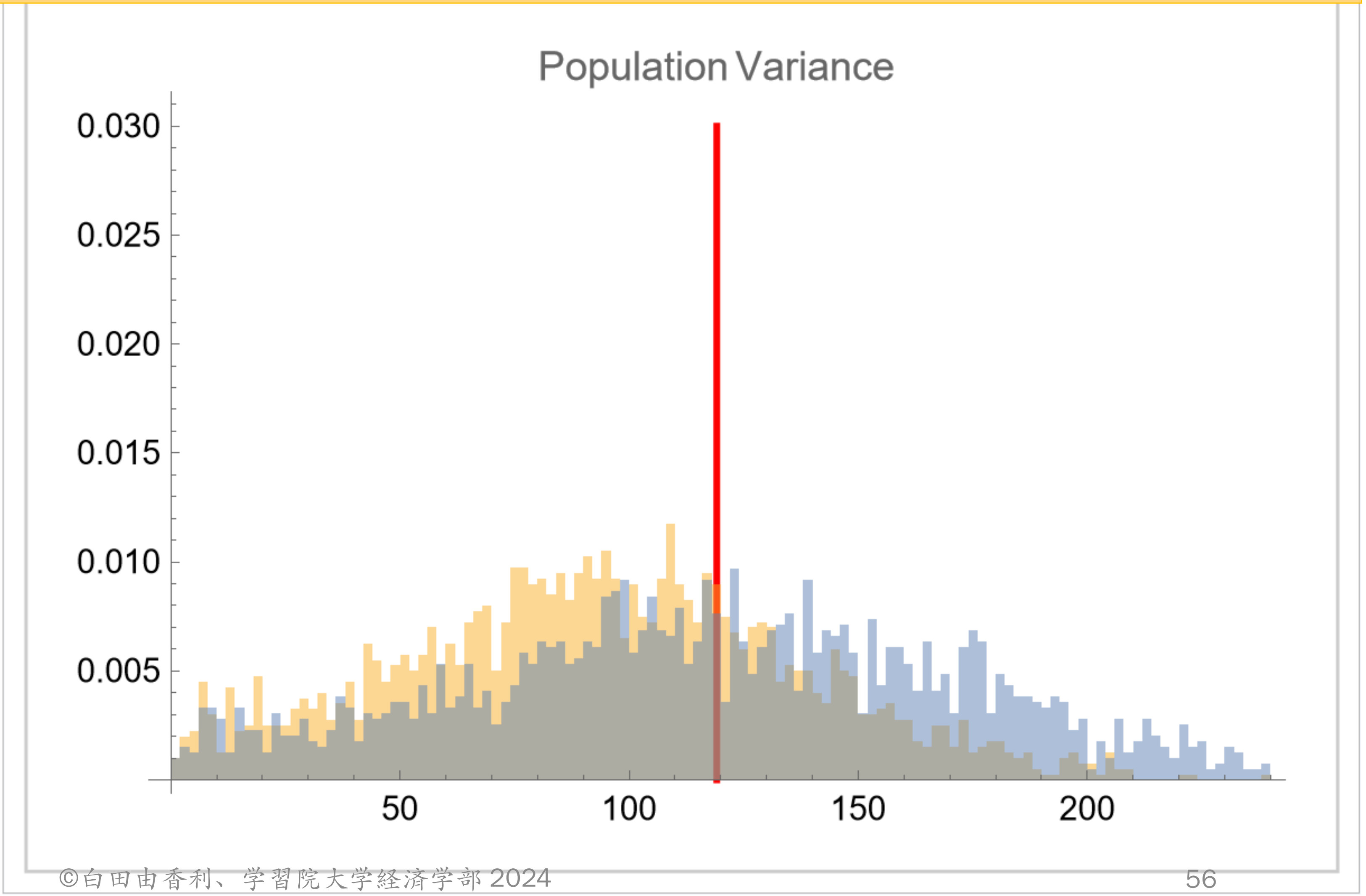
不偏分散

- 標本から母分散を推測したい.

2000回標本取ってきた

- 青の分散のほうが不偏分散で計算（n -1)
- 黄色の分散はnで割っている
- 赤の値は母分散
- どちらが母分散の周りに集まっていますか？

グラフィクス教材
www-cc.gakushuin.ac.jp/~20010570/VDStat/



©白田由香利、学習院大学経済学部 2024