シミュレーション論 Ⅱ

第１３回

データの分類と数値処理（２）

第12回のレポート

- ４つの駅で、一日の乗降人数と売店の売上を調べたところ、以下の表のようになった。乗降人数と売上の関係を最小二乗法を用いて線形回帰せよ。
- あらたに乗降人数が5千人の駅に売店を出すことにした。いくらの売上が見込まれるか推定せよ。

駅	乗降人数(千人)	売上(万円)
A駅	1	19
B駅	2	23
C駅	3	35
D駅	4	37

第12回のレポート解答(1)

- データ数 n = 4 で、データの組はそれぞれ

$$(x_1, y_1) = (1, 19), \quad (x_2, y_2) = (2, 23), \quad (x_3, y_3) = (3, 35), \quad (x_4, y_4) = (4, 37)$$

a, b は以下の式で求められる

$$\boxed{\begin{cases} a = \dfrac{nD - CE}{nB - E^2} \\ b = \dfrac{BC - DE}{nB - E^2} \end{cases}}$$

- ただし

$$B = \sum_{i=1}^{n} x_i^2, \quad C = \sum_{i=1}^{n} y_i$$

$$D = \sum_{i=1}^{n} x_i y_i, \quad E = \sum_{i=1}^{n} x_i$$

第12回のレポート解答(2)

- B～Eを求めると

$$B = \sum_{i=1}^{n} x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30$$

$$C = \sum_{i=1}^{n} y_i = 19 + 23 + 35 + 37 = 114$$

$$D = \sum_{i=1}^{n} x_i y_i = 1 \times 19 + 2 \times 23 + 3 \times 35 + 4 \times 37 = 318$$

$$E = \sum_{i=1}^{n} x_i = 1 + 2 + 3 + 4 = 10$$

第12回のレポート解答(3)

- a, b の式に代入して

$$a = \frac{nD - CE}{nB - E^2} = \frac{4 \times 318 - 114 \times 10}{4 \times 30 - 10^2} = 6.6$$

$$b = \frac{BC - DE}{nB - E^2} = \frac{30 \times 114 - 318 \times 10}{4 \times 30 - 10^2} = 12$$

- よって売上の推定値(回帰直線)は

  Y = 6.6X + 12

- 5000人の乗降客がある駅に新たに店を出した場合の売り上げ予想は

  6.6×5 + 12 ＝ 45 (万円)

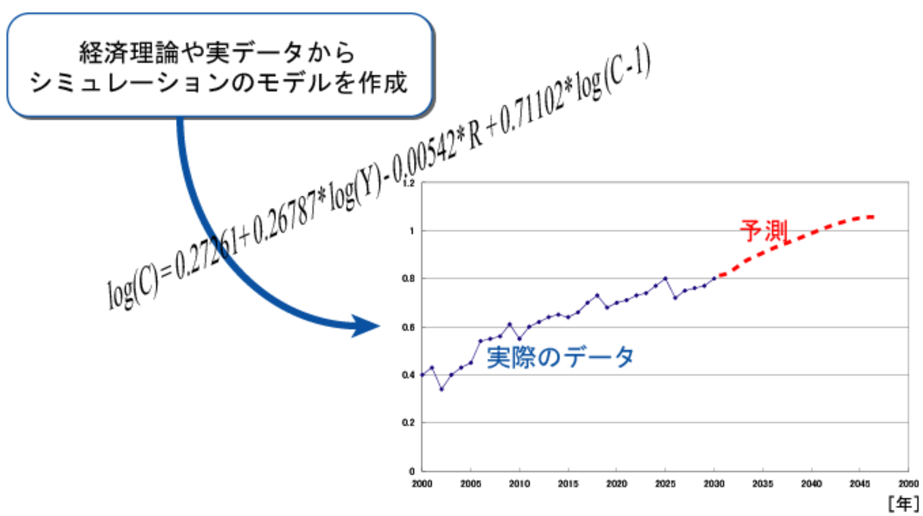
今回の内容

- ２つの変数間に線形の関係がある場合→線形回帰
- 線形回帰をおこなうにあたって、気になること
  変数間の関係性はどのようなものか？
  そもそも本当に線形の関係があるのか？
- 変数の間の関係性（相関）を分析する

データの相互関係

- 実験やシミュレーション結果、実際の観察データ、種々の統計データなどからデータ間の関係を推定・整理する
- 原因-結果の相互関係が明確になり、モデルの作成や将来の予測が可能となる

→回帰分析・相関分析など



線形回帰の注意点

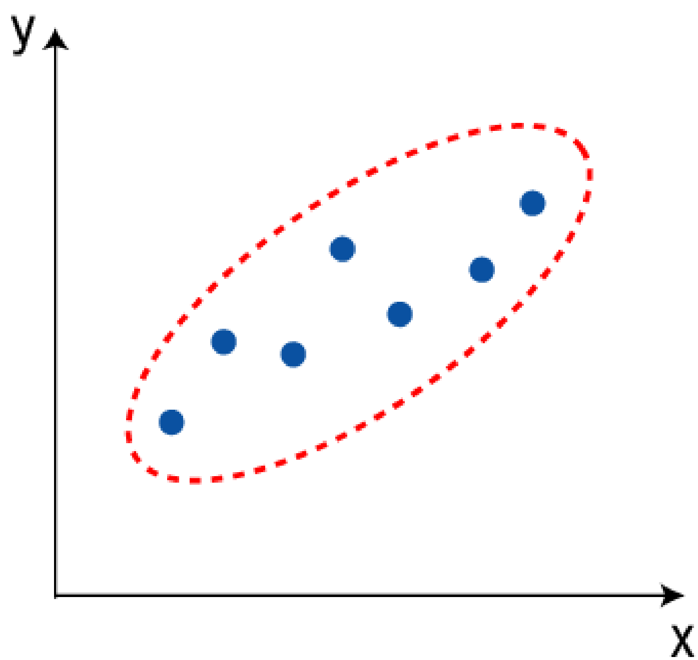
- なんでもかんでも線形回帰をおこなえばいいというわけではない
- 線形回帰は2種類のデータ(2変数)間に「直線的な関係がある」ことが前提になる
- データによって関係性が強いもの、弱いものがある
  →2変数の関係(相関)の強さを調べてやる必要がある
- 相関係数の導出

相関係数

- 2変数に関連性がある（片方が大きくなると、もう一方も大きくなる、など）場合、2変数に「相関」があるという
- 相関係数：相関の度合いを表す数値で、Rで表す
- 相関係数は－１～＋１の数値をとる
  - **片方が増えればもう片方も増える→相関係数は＋、正の相関**
  - **片方が増えればもう片方が減る　→相関係数は－、負の相関**
- 絶対値が大きいほど、相関が強い（2変数の間に強い関連性がある）

相関係数(正の相関)

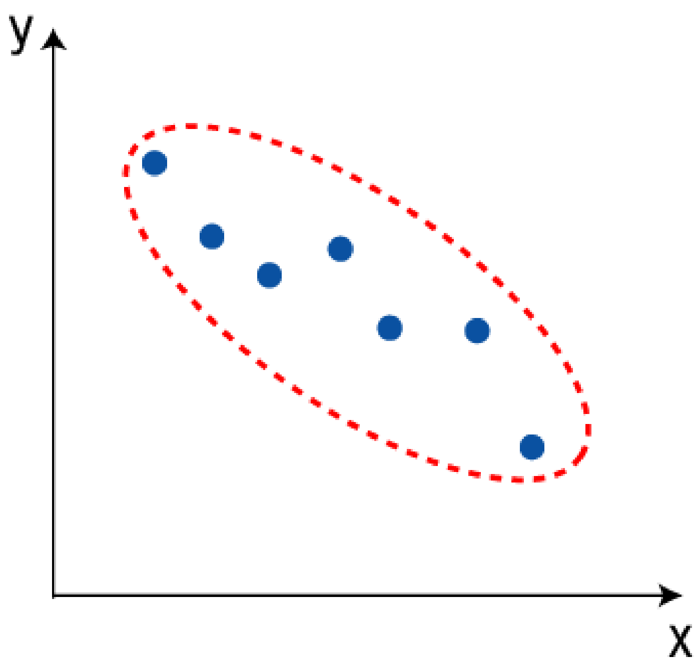
- 2変数に「片方が増えれば、もう片方も増える」という関係がある場合を「正の相関」という



相関係数 $R>0$

相関係数(負の相関)

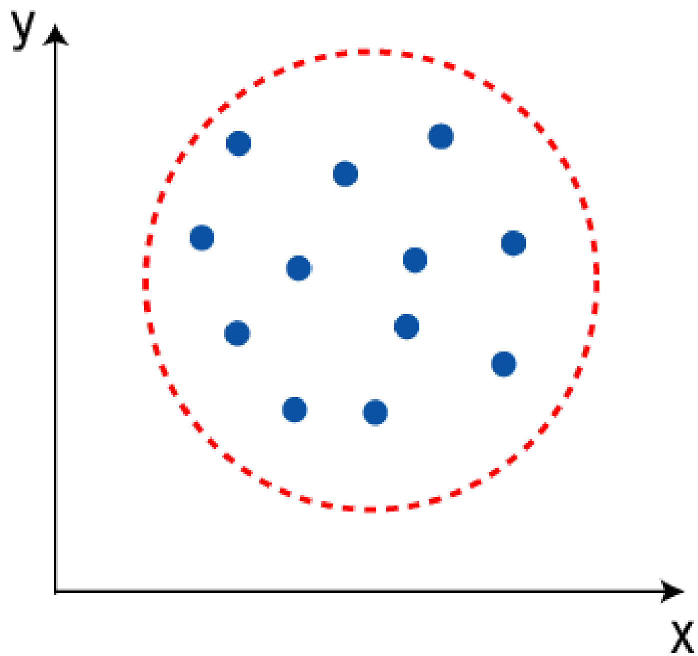
- 2変数に「片方が増えれば、もう片方が減る」という関係がある場合を「負の相関」という



相関係数　$R < 0$

相関係数(無相関)

- 2変数に(線形の)関係が(ほとんど)見られない場合、データは円に近い形で分布する。このような場合を無相関という
- このとき相関係数は0に(近く)なる



相関係数 $R \fallingdotseq 0$

相関係数の式

- n 組のデータ　(x_1, y_1),　(x_2, y_2),　$(x_3, y_3)\Lambda$　(x_n, y_n)
  があるとき、x, y の相関係数 R は

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- ただし　$\bar{x}, \bar{y}$　はそれぞれ x, y の平均 $\frac{1}{n}\sum_{i=1}^{n} x_i$, $\frac{1}{n}\sum_{i=1}^{n} y_i$

相関係数の値と関係性の度合い

- 相関係数の絶対値が 1 に近いほど相関の度合いは強くなる
- 一般的には以下のような分類になる
- 相関係数の絶対値が1の場合を「完全な相関」といい、データが一直線上に並ぶ場合に相当する

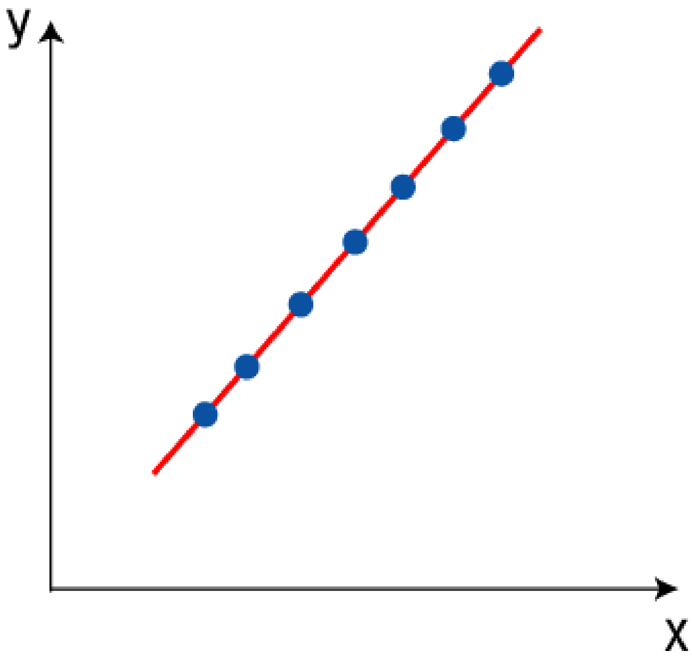
Rの値	相関の度合い
0	無相関
0<|R|≦0.2	ほとんど相関なし
0.2<|R|≦0.4	低い相関あり
0.4<|R|≦0.7	相関あり
0.7<|R|<1.0	高い相関あり
1.0 または−1.0	完全な相関

注:これ以外の分類方法もある

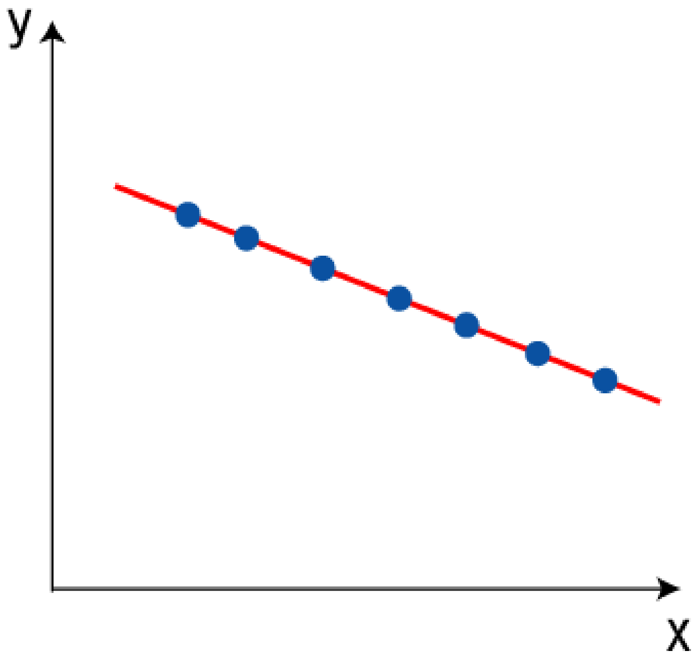
相関係数の値とデータの並び

- 相関係数 R の絶対値が1のときは以下のようにデータが一直線上に並ぶ

完全な正の相関
（R ＝ 1）



完全な負の相関
（R ＝ −1）



相関係数の計算(1)

- 以下のようなデータがあるとき、相関係数を計算してみよう

$$(x_1, y_1) = (1, 19)$$
$$(x_2, y_2) = (2, 23)$$
$$(x_3, y_3) = (3, 35)$$
$$(x_4, y_4) = (4, 37)$$

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$x\text{の平均値}：\bar{x} = \frac{1+2+3+4}{4} = 2.5$$

$$y\text{の平均値}：\bar{y} = \frac{19+23+35+37}{4} = 28.5$$

相関係数の計算(2)

- Rの分子を計算する

$$
\begin{aligned}
\text{分子}:\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_i-2.5)(y_i-28.5) \\
&=(1-2.5)(19-28.5)+(2-2.5)(23-28.5) \\
&\quad +(3-2.5)(35-28.5)+(4-2.5)(37-28.5) \\
&=14.25+2.75+3.25+12.75 \\
&=33
\end{aligned}
$$

相関係数の計算(3)

- Rの分母を計算する

$$
\text{分母}: \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}
$$

$$
= \sqrt{\frac{\{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2\}}{\times\{(19-28.5)^2+(23-28.5)^2+(35-28.5)^2+(37-28.5)^2\}}}
$$

$$
= \sqrt{(2.25+0.25+0.25+2.25)\times(90.25+30.25+42.25+72.25)}
$$

$$
= \sqrt{5\times 235}
$$

$$
= 34.278\ldots
$$

相関係数の計算(4)

- Rの値を計算する

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{33}{34.278} = \boxed{0.962\ldots}$$

- 参考：Excelでは関数 CORREL(xの範囲, yの範囲)で計算できる

	A	B
1	x	y
2	1	19
3	2	23
4	3	35
5	4	37
6		
7	相関係数R	0.96270894
8		

=CORREL(A2:A5,B2:B5)

相関係数の検定（無相関検定）

- $R = 0.962$ だから、一般的には「x, y の間に高い相関がある」と言える
- しかし、相関係数の値はデータの個数やバラツキに大きく左右される

  →「たまたま」そうなったのか、関係性があるからそうなったのか
- 2変数の間にちゃんと「相関があるかどうか」を調べてやる必要がある

  →無相関検定

無相関検定

- 2変数の間に「相関がない」と仮定して(帰無仮説)、有意性の検定をおこなう
- 相関係数に意味があるかどうか(相関があるかどうか)
- 以下の表で、5%有意水準の値より相関係数 R が大きければ「有意」、つまり相関があるといえる(1%水準はさらに厳しい検定)

(実際には帰無仮説の棄却なので「無相関であるとはいえない」ということ)

データ数 (n)	5%有意水準	1%有意水準
3	0.997	1.000
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
100	0.197	0.256
1000	0.062	0.081

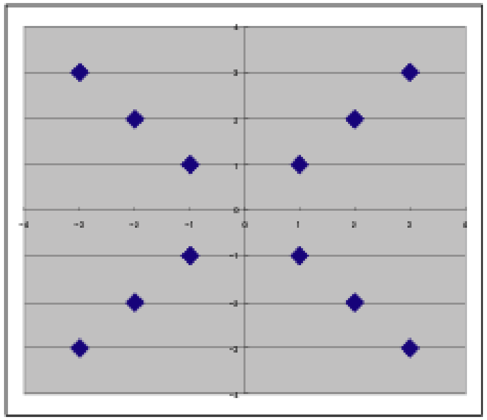
無相関検定（2）

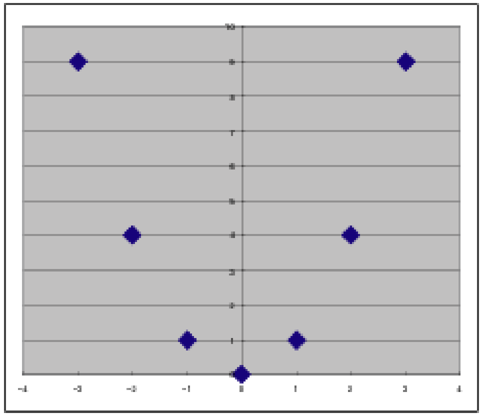
- 先ほどの練習問題だと $n = 4$ だったから、
  $R > 0.950$ で 5%有意、$R > 0.990$ で 1%有意となる。
- $R = 0.962$ だったので5%有意となり「2変数 x, y に相関がある」といえる

データ数（n）	5%有意水準	1%有意水準
3	0.997	1.000
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
100	0.197	0.256
1000	0.062	0.081

相関分析の注意点

- 相関係数の絶対値が1に近くても必ずしも相関があるとはいえない（無相関検定）
- 相関分析はあくまで「相関がありそう」という判定
  →まったく無関係のデータ間に強い相関が見られることもある
- ここまで述べてきた相関係数は実は「ピアソンの積率相関係数」というもので、2変数間に線形の関係があることを前提にしたもの
- データが曲線に沿う場合や、他の関係がある場合には意味をなさない





例：どちらも相関係数（ピアソンの積率相関係数）は0だが、
果たして変数間に関係性がないといえるだろうか？

相関分析の注意点(2)

- 変数間の関係性を調べたいときには、まず散布図を描いてみる
- データの散らばりからある程度の仮説を立て、それに応じた分析方法を使う
- 線形の関係性が見られる場合→相関分析、線形回帰分析
- 適切なデータの分析法を使用して
  ・ 実験データ、統計データからシミュレーションモデルの構築
  ・ シミュレーション結果の整理と分類
  などに使用することができる

第13回のレポート

- 気温とおでんの売り上げが以下の表のようになった。

（1）相関分析をおこない、気温とおでんの売り上げに相関があるかどうか、あるならばどのような相関か調べよ
  - **相関係数を計算し、無相関検定により有意性を判定する**
  - **相関係数の値から相関の正負、強弱を調べる**

（2）線形回帰をおこない、回帰直線の式を求めよ

注：無相関検定は R の絶対値で判定してよい

気温（℃）	売り上げ（個）
18	
15	
10	
5	