

MAT220 Lecture 18: Linear Correlation (線形相関)

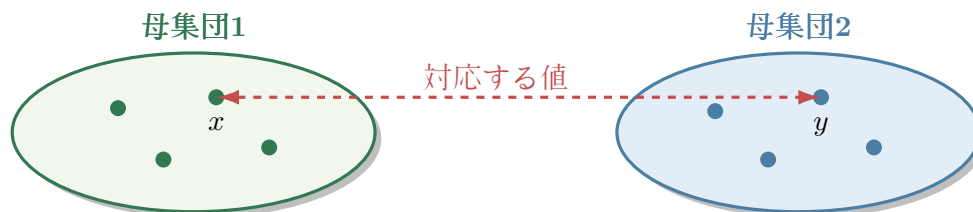
Contents

1	相関とは何か	2
1.1	散布図	2
1.2	線形相関	3
1.3	相関の強さと符号	4
1.4	符号と強さ	5
2	相関係数	6
2.1	標本相関係数	6
2.2	r の公式	7
2.3	共分散を理解する	8
2.4	r の公式を理解する	11
2.5	完全な正の相関から $r = 1$ を求める	11
2.6	例: r の計算	13
3	相関の限界	14
3.1	r が定義されない場合	14
3.2	非線形相関	15
3.3	因果関係を伴わない相関	15

前の二つの講義では、二つの母集団について同時に仮説検定を行いました。この講義では、対応のあるデータをまったく異なる方法で使います。母数を互いに比較して検定する代わりに、対応のあるデータの記述統計を調べます。ここでは、二つのデータ集合が直線に似たパターンを示すかどうかを考える線形相関を扱います。まず直感的に、つまり視覚的に考え、その後で線形関係を数学的に表す方法を調べます。

1 相関とは何か

第16回講義では、対応のある母集団について説明しました。簡単に言えば、一方の母集団の各値に対して、もう一方の母集団に自然な対応相手が存在する母集団です。



対応のあるデータは、いくつかの方法で得られます。

- 同じ集団に対する連続した実験：例えば、カフェイン摂取前と摂取後の反応時間。
- 同じ対象に対する異なる測定：例えば、同じランニング中に胸部ストラップとApple Watchで測定した心拍数。
- 自然に対称または対応している対象：例えば、一卵性双生児の第一子と第二子、または同じ人の左足と右足。

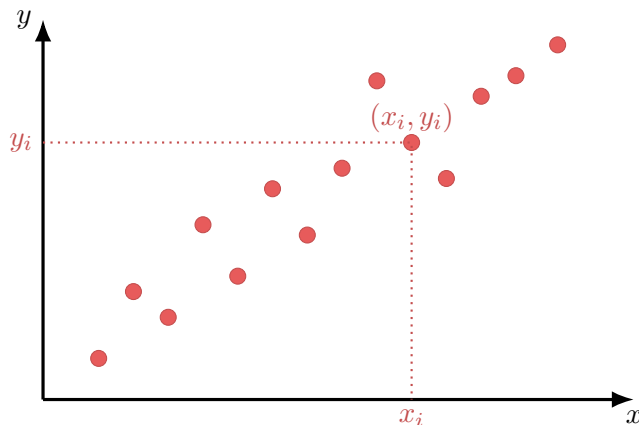
以前は、対応関係を利用して確率変数の差を作り、 $D = X - Y$ と書きました。そして、 X と Y の平均が異なるかどうかを検定しました。今回は、対応のあるデータの記述だけを調べます。まず、散布図と呼ばれる視覚的な表現から始めます。

1.1 散布図

n 組のデータから標本を取るとします。これは次のリストとして書けます。

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

散布図とは、各組 (x_i, y_i) を二次元空間内の一つの点として描いた、データの視覚的表現です。横軸は x 変数を、縦軸は y 変数を表し、各点が一組のデータを表します。



実際には、この種類のデータは、同じ対象を二つの異なる方法で測定するときに得られます。対応する観測値を描くことによって、データに何らかのパターンがあるかどうかを視覚的に調べることができます。数学的な設定では、変数 Y を変数 X によって説明しようとするのがよくあります。このため、 x を説明変数、 y を応答変数と呼びます。¹

散布図

散布図とは、対応のある量的データの各組 (x_i, y_i) を一つの点として描いたグラフです。変数 x を説明変数、変数 y を応答変数と呼びます。

この用語は、私たちが考えている問題の種類を示しています。 x の変化が、なぜ y が変化するかを説明する助けになると考えます。例えば、 x を車の速度、 y を停止距離とします。速度が停止距離を説明するのに役立ち、停止距離が何らかの規則に従って速度の変化に応じて変化すると考えるのは自然です。

1.2 線形相関

「相関」とは、二つの変数と一緒に何らかのパターンを示すことを意味します。原理的には、二つの変数の間のパターンには多くの異なる形があり、二つのデータ集合はさまざまな方法で相関する可能性があります。この講義では、最も単純な場合である線形相関を扱います。大まかに言えば、点のパターンが直線にほぼ従うとき、データには線形相関があります。

線形相関

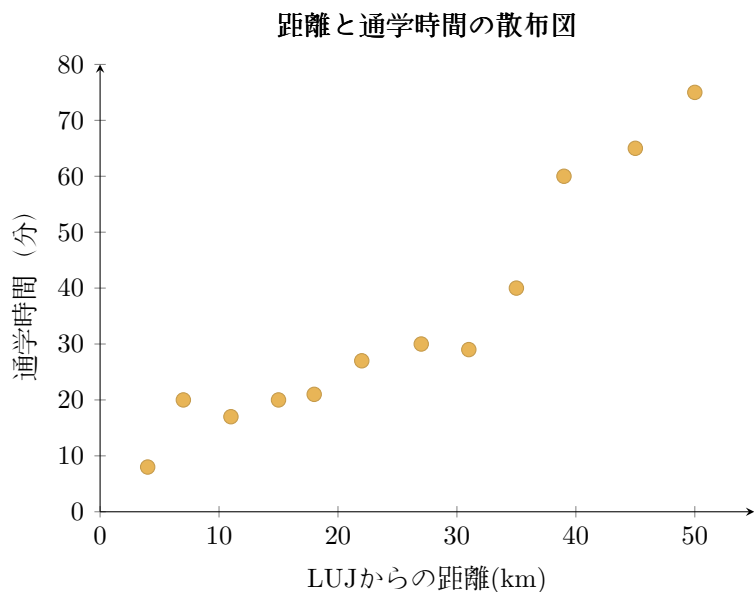
散布図に直線のように見える全体的なパターンがあるとき、対応のあるデータは**線形相関**を示します。

線形関係は非常によく見られます。例えば、地域の人口とGDP、ばねを伸ばすために必要な力、酸素消費量とエネルギー消費量などがあります。第19回講義では、線形パターンを最もよく近似する直線を求める方法である線形回帰を導入します。

例として、12人の学生を調査し、自宅から大学までの距離と、大学までの通学時間を記録したとします。このデータには自然な対応があります。二つの値が同じ学生に関係しているからです。データと対応する散布図を以下に示します。

¹文献では、この二つの変数に別の名称が使われることもあります。 x を独立変数、 y を従属変数と呼ぶ場合があります。

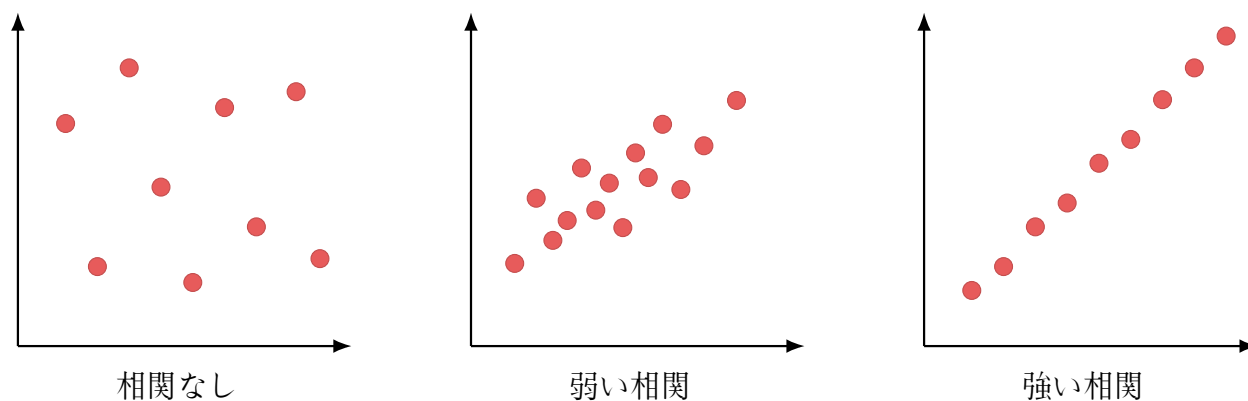
学生	距離 (km)	時間 (分)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75



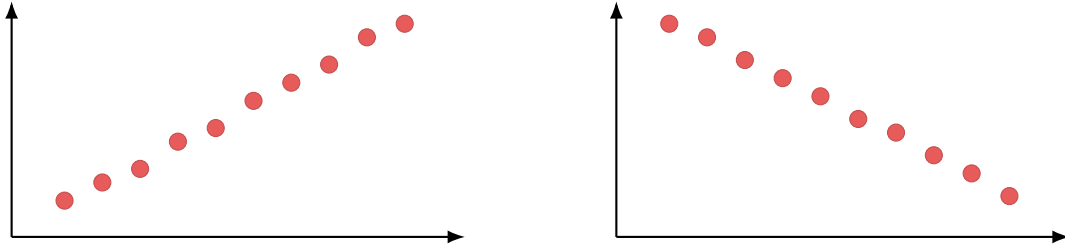
散布図を見ると、データ点には明確な線形パターンがあります。大学の近くに住む学生は通学時間が短く、大学から遠くに住む学生は通学時間が長くなっています。もちろん、このデータは完全な直線上には並びません。これは当然のことです。鉄道路線、移動手段など、通学時間に影響する小さな要因が多数存在するからです。それでも、データから全体的なパターンを見ることができます。

1.3 相関の強さと符号

一般に、明確なパターンがまったくない散布図もあれば、直線に非常によく似た散布図もあります。通常の言葉では、この違いを相関の「強さ」で表します。大まかに言えば、パターンが非常に明確なとき相関を「強い」と呼び、パターンが多少は存在するものの、あまり明確でないとき相関を「弱い」と呼びます。



これに加えて、線形相関にはもう一つ重要な特徴があります。次の二つの図を考えます。



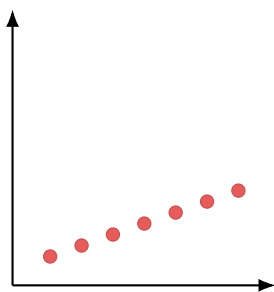
どちらも強い線形相関を示していますが、明らかに異なっています。左のグラフでは、 x と y は一緒に増加します。 x が小さいとき y も小さく、 x が大きいとき y も大きくなります。一方、右のグラフでは関係が逆です。 x が小さいとき y は大きく、 x が大きいとき y は小さくなります。つまり、これらのデータ集合に当てはまる直線の傾きは、異なる符号を持ちます。このことから、次の用語を導入します。

相関の符号

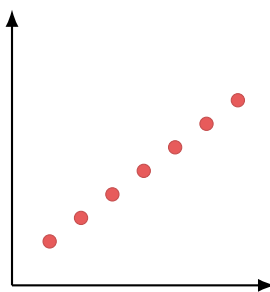
相関の符号は、全体的な傾向の方向を表します。

- **正の相関**： x と y は一緒に増加する傾向があり、全体的な傾向は正の傾きを持ちます。
- **負の相関**： x が増加すると y は減少する傾向があり、全体的な傾向は負の傾きを持ちます。

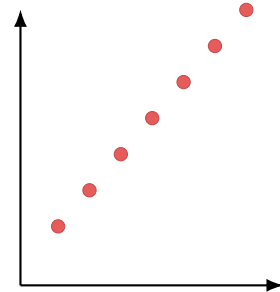
ここで、「正」と「負」という言葉は、パターンの符号だけを表し、傾きの大きさを表すものではありません。例えば、次の三つのグラフは傾きが異なりますが、すべて完全な正の線形相関を示します。



完全な正の相関



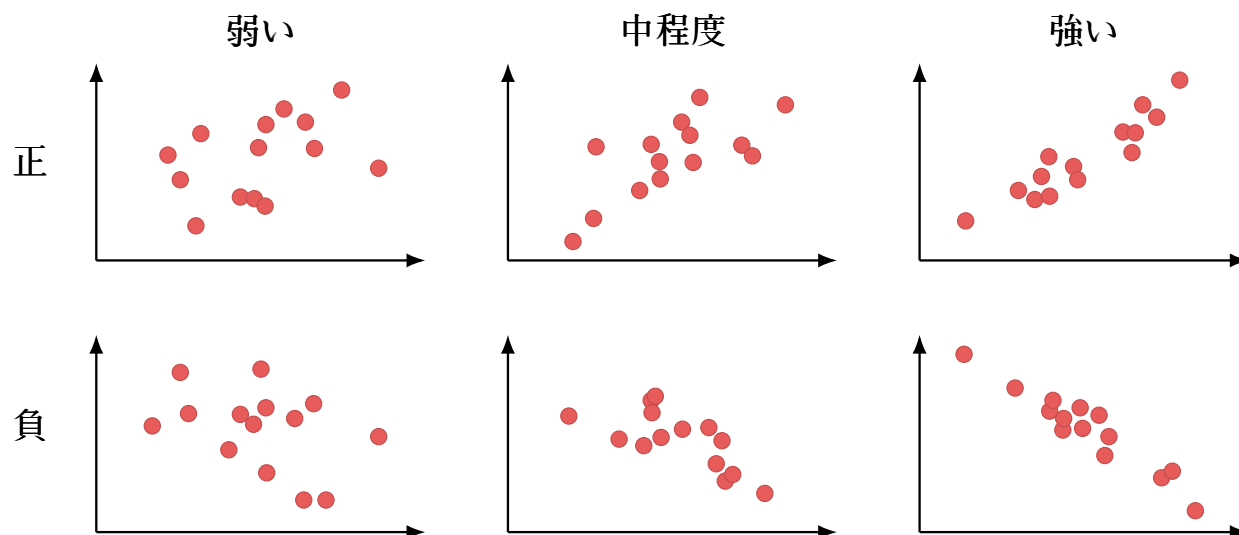
完全な正の相関



完全な正の相関

1.4 符号と強さ

符号と強さは、線形相関の二つの異なる特徴です。次の図が示すように、正と負のどちらの符号についても、弱い、中程度、強いパターンがあります。



2 相関係数

2.1 標本相関係数

「弱い」や「強い」のような自然言語による表現は直感的ですが、やや曖昧です。この問題を解決するために、相関の強さを数値で表す量を導入します。簡単に言えば、相関が強いほど絶対値が大きく、相関が弱いほど絶対値が小さくなる数を定義します。この数は通常、標本相関係数と呼ばれ、 r と書きます。標本相関係数は、 x と y の間の線形関係の強さと符号の両方を測るように作られています。

標本相関係数

標本相関係数 r は単位を持たない数であり、

$$-1 \leq r \leq 1$$

を満たします。

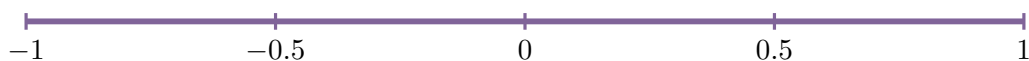
r の符号は線形相関が正か負かを表し、 r の絶対値は線形関係の強さを表します。

「標本相関係数」という用語は、三つの部分に分けて考えることができます。

- 「係数」は、 r が数であることを表します。
- 「相関」は、 x と y の間のパターンを表します。
- 「標本」は、 r が標本データ (x_i, y_i) から作られることを表します。

したがって、「標本相関係数」とは、簡単に言えば「標本がどの程度直線のパターンに従うかを測る数」です。 r を使うことで、線形相関の強さをより正確に表すことができます。

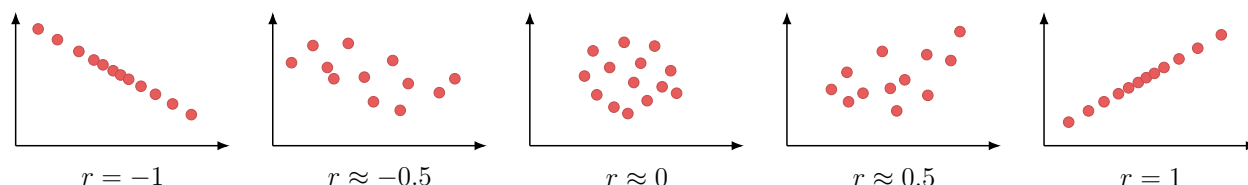
完全な 強い 弱い 線形相関 弱い 強い 完全な
負の相関 負の相関 負の相関 なし 正の相関 正の相関 正の相関



重要な基準値を言葉で表すと、次のようになります。

- $r = 1$ は、データが正の傾きを持つ直線上に完全に並んでいることを意味します。
- $0 < r < 1$ は、正の線形相関があることを表します。
- $r = 0$ は、データに線形相関がまったくないことを表します。
- $-1 < r < 0$ は、データに負の線形相関があることを表します。
- $r = -1$ は、完全な負の線形相関を表します。

次の図は、 r のいくつかの異なる近似値を持つデータを示しています。



x と y を結び付ける正確な数学的規則が存在しても、現実のデータがその規則に完全に従うことはほとんどありません。測定にはノイズ、個人差、小さな誤差が含まれます。したがって、完全な線形相関は理想化された理論上の極限と考えるのが適切です。

この講義では直接扱いませんが、線形相関を測る母数も存在します。平均や標準偏差と同様に、母数はギリシャ文字で表します。相関の場合は文字 ρ を使います。

	標本	母集団
平均	$\bar{x}$	μ
標準偏差	s	σ
相関係数	r	ρ

標本相関係数 r は、与えられた標本内の点の組の間にある線形関係を測ります。一方、母相関係数 ρ は、母集団全体のすべての組の間にある真の相関を表します。

2.2 r の公式

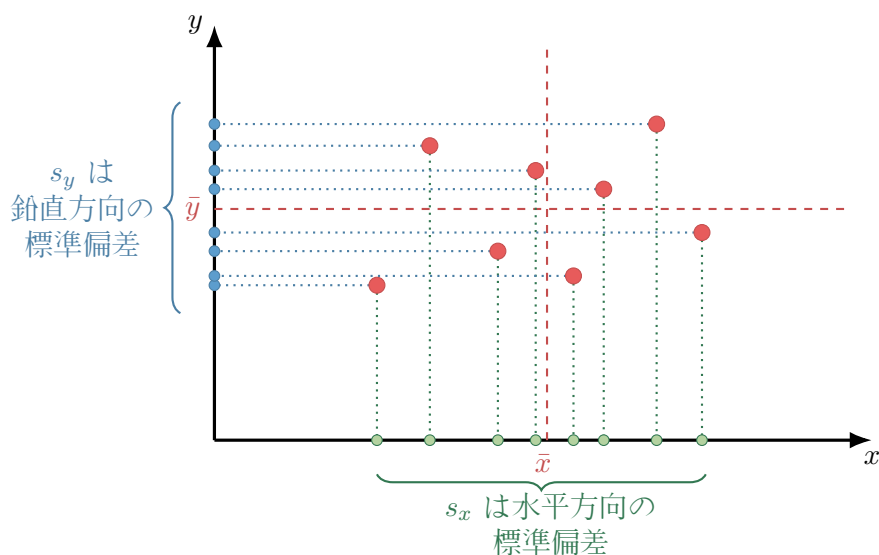
標本相関係数は次の公式で定義されます。

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

一見すると、この公式は非常に複雑に見えます。何が起きているのかを理解するために、公式を段階的に分解します。まず、 r の分母には二つの標本標準偏差があることに注意してください。

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{および} \quad s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

ここでは対応のあるデータを扱っています。つまり、標本内の各データ点は (x_i, y_i) の形をしています。それぞれの組を二つの座標に分け、個別に考えることができます。



したがって、 r の公式では、 x と y のそれぞれの広がりで割っています。 r の分子には、共分散と呼ばれる新しい項があります。

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

実際、これは第16回講義で対応のある標本の分散を説明したときに触れた共分散の標本版です。そのときは、 x と y の相互関係を測る量として、その存在を述べただけでした。ここでは、その意味を詳しく説明します。

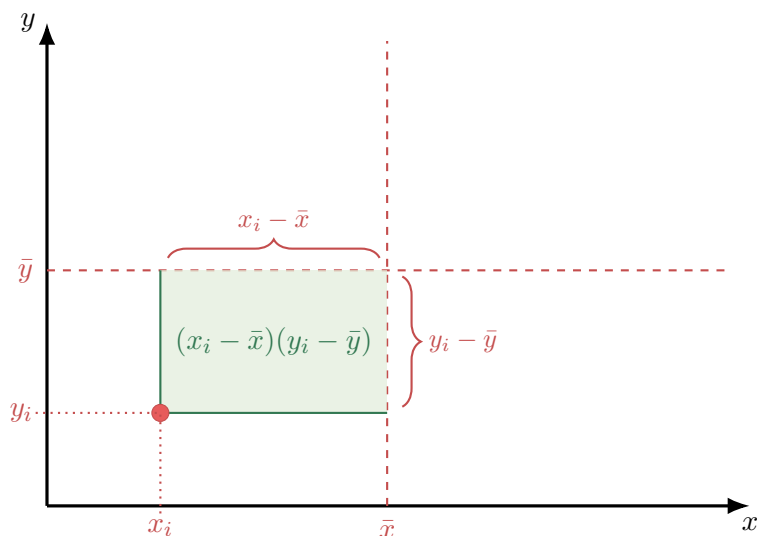
2.3 共分散を理解する

共分散の公式を調べます。

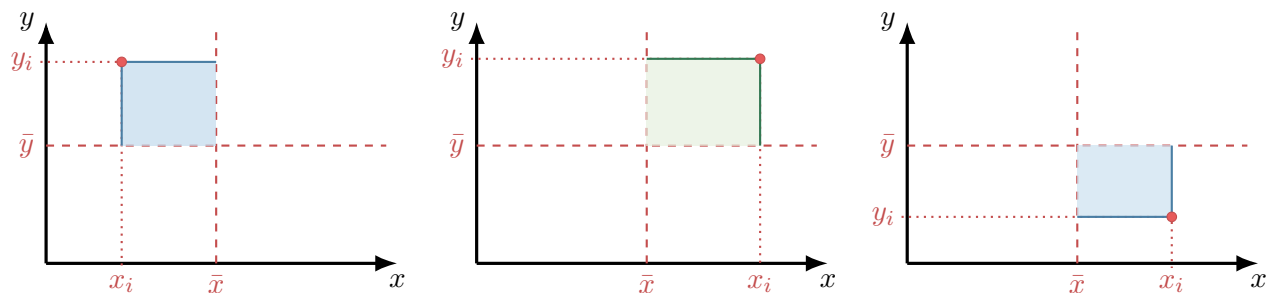
$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

この公式では何が起きているのでしょうか。まず、外側では多くの量を合計し、その後で $n-1$ で割っています。この部分は標本分散の公式に似ています。実際、一種の平均のような形で

す。それでは、合計している各量は何でしょうか。 $(x_i - \bar{x})$ と $(y_i - \bar{y})$ は、点と平均の間の距離を測っています。具体的には、 x_i が $\bar{x}$ からどれだけ離れているか、 y_i が $\bar{y}$ からどれだけ離れているかを測ります。これらは二つの長さであり、それらを掛けると面積ができます。



上の図では、データ点 (x_i, y_i) を意図的にグラフの左下の領域に描いています。この場合、 x_i は $\bar{x}$ の左側にあるため、 $x_i < \bar{x}$ であり、したがって $x_i - \bar{x} < 0$ です。同様に、 y_i は平均 $\bar{y}$ の下側にあるため、 $y_i < \bar{y}$ であり、したがって $y_i - \bar{y} < 0$ です。したがって、 $(x_i - \bar{x})$ と $(y_i - \bar{y})$ は符号を持つ距離です。² これらの符号付き距離が両方とも負であれば、その積は正になります。他の三つの領域では、面積の符号は次のようになります。



各長方形の符号は、点が二つの平均のどちら側にあるかで決まります。

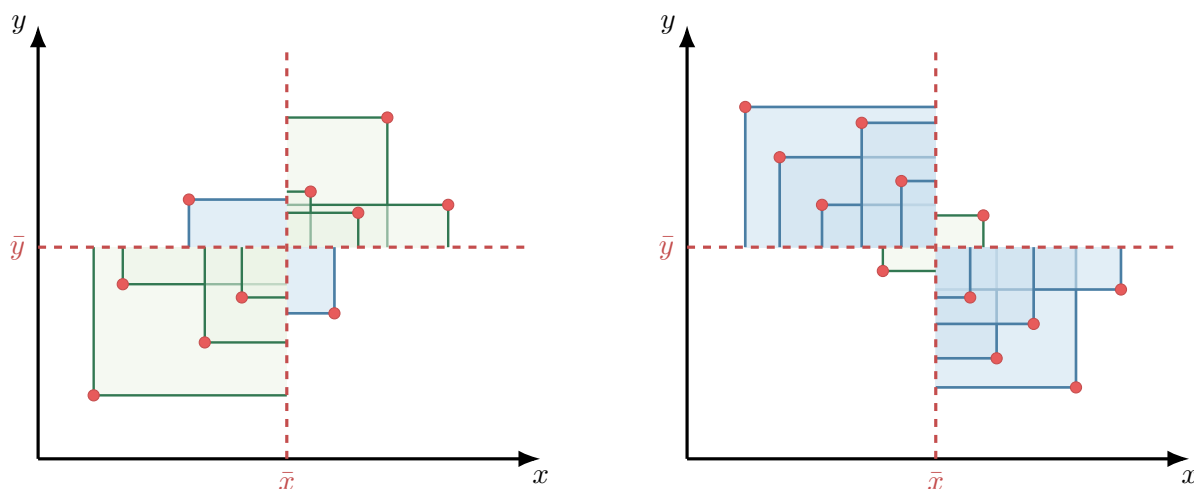
点 (x_i, y_i) の位置	$(x_i - \bar{x})(y_i - \bar{y})$	符号付き面積
$x_i < \bar{x}, y_i < \bar{y}$	$(-)(-)$	正
$x_i < \bar{x}, y_i > \bar{y}$	$(-)(+)$	負
$x_i > \bar{x}, y_i > \bar{y}$	$(+)(+)$	正
$x_i > \bar{x}, y_i < \bar{y}$	$(+)(-)$	負

$x_i = \bar{x}$ または $y_i = \bar{y}$ であれば、符号付き面積は0 です。

² これらは平均からの符号付き偏差と呼ばれることがあります。

共分散は、これらすべての長方形の符号付き面積を合計し、その後で $n - 1$ で割ります。つまり、共分散は、二つの変数がそれぞれの平均の同じ側にあるかどうか、また平均からどの程度離れているかを平均的に測ります。

実際、この「符号付き面積の一種の平均」によって、相関の符号を求めることができます。次の図では、各データ点について符号付き長方形を描いています。正の符号を持つ面積を緑色で、負の符号を持つ面積を青色で表します。



左の図には、ある程度の正の相関があります。データの多くが図の左下または右上の領域にあります。視覚的には、青色より緑色の面積の方が大きいため、共分散は正になるはずですが。一方、右の図には負の相関があります。データ点の多くが図の左上または右下にあります。符号付き長方形の大部分が青色であるため、共分散は負になります。まとめると、次のようになります。

- 図に青色より緑色の面積が多ければ、符号付き面積の合計は正であり、 $\text{cov}(x, y) > 0$ です。
- 図に緑色より青色の面積が多ければ、符号付き面積の合計は負であり、 $\text{cov}(x, y) < 0$ です。
- 正と負の面積が合計の中でほぼ打ち消し合えば、共分散は0に近くなります。

公式に戻ると、それぞれの記号は次のように分解できます。

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

点と $\bar{x}$ の間の
水平方向の符号付き距離

点と $\bar{y}$ の間の
鉛直方向の符号付き距離

すべての面積を平均

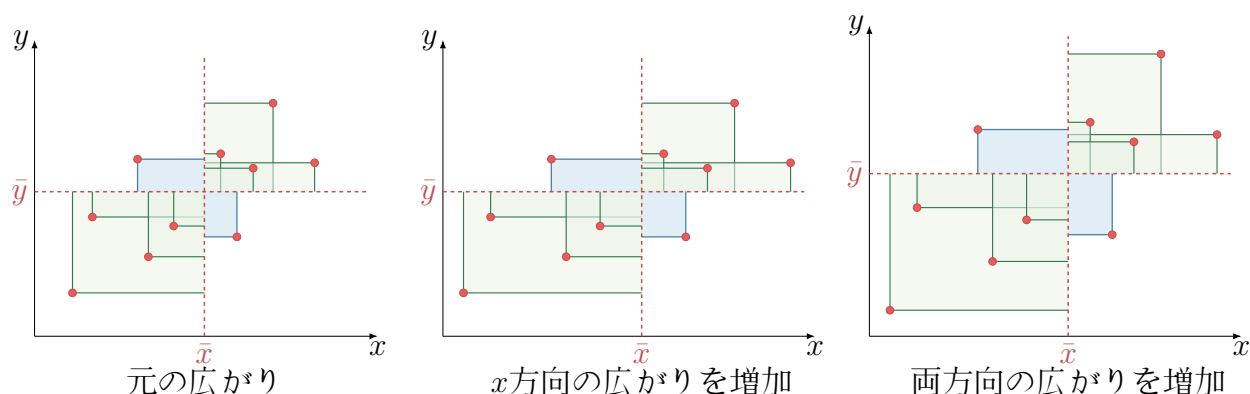
長方形の符号付き面積

共分散が測るもの

共分散は、二つの変数がそれぞれの平均の同じ側にある傾向があるかどうかを測ります。

2.4 r の公式を理解する

共分散の公式は、長方形の面積を利用して、線形相関の符号を記録します。しかし、共分散だけでは、依然としてデータ全体の広がり依存します。これを確認するために、一つのデータ集合について、 x 方向と y 方向の広がりを手動で増やしてみます。



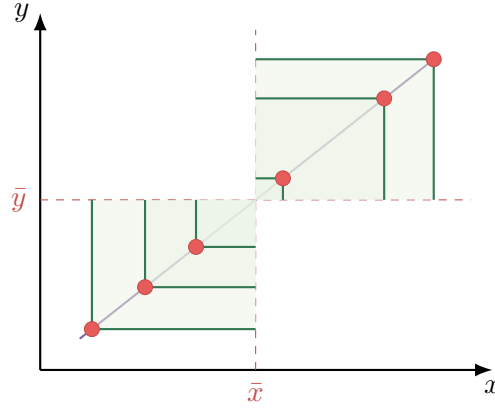
データの広がりが増やすと、各点から $\bar{x}$ または $\bar{y}$ までの平均距離も増加するため、各長方形の平均面積も増加します。しかし一般に、この方法で広がりが増やしても、緑色の長方形が青色に変わったり、青色の長方形が緑色に変わったりすることはありません。長方形の実際の大きさではなく、緑色と青色の長方形の相互関係だけを考えたいなら、データの広がりですべての絶対的な大きさを取り除くことができます。これが r の公式です。

$$r = \frac{\text{cov}(x, y)}{s_x s_y}.$$

大まかに言えば、標準偏差 s_x と s_y は、それぞれ水平方向と鉛直方向におけるデータの広がり測ります。共分散をこれら二つの項で割ると、上の図の全体的な尺度が取り除かれるため、データの広がりが増やしても答えは変わりません。その代わりに、データから作られる面積の相対的なパターンだけが重要になります。

2.5 完全な正の相関から $r = 1$ を求める

データが完全に直線上に並んでいる場合を考えます。 r の公式が $+1$ を与えることを示します。次の図では、正の傾きを持つ直線上に六つの点を描き、それぞれの符号付き面積を示しています。



幾何学的には、これらの長方形はすべて同じ縦横比を持ちます。

$$\bar{y} - y_1 \left\{ \underbrace{\hspace{2cm}}_{\bar{x} - x_1} \right\} \sim \underbrace{\hspace{1cm}}_{\bar{x} - x_2} \bar{y} - y_2 \Rightarrow \frac{y_1 - \bar{y}}{x_1 - \bar{x}} = \frac{y_2 - \bar{y}}{x_2 - \bar{x}}$$

この縦横比を k と呼びます。すべての点 (x_i, y_i) が同じ直線上にあるなら、それぞれの比は k に等しくなります。³式を変形すると、すべての $i = 1, \dots, n$ について、

$$\frac{y_i - \bar{y}}{x_i - \bar{x}} = k \Rightarrow y_i - \bar{y} = k(x_i - \bar{x})$$

となります。

これを共分散の公式に代入すると、 x の標本分散を含む式が得られます。

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})k(x_i - \bar{x}) = \frac{k}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = ks_x^2.$$

同様に、 s_y の公式に含まれるすべての $y_i - \bar{y}$ を $k(x_i - \bar{x})$ に置き換えると、

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

となります。最後の等号では $k > 0$ を使っています。これらを r の公式に代入すると、

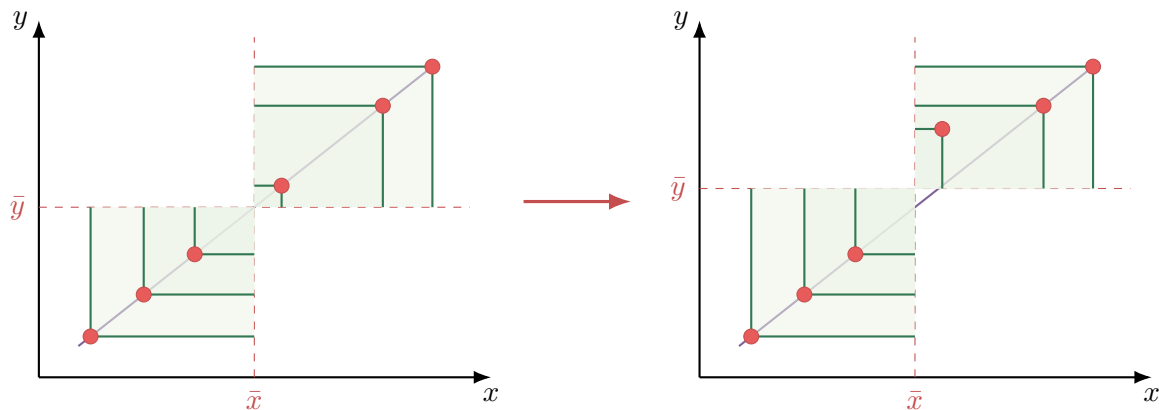
$$r = \frac{ks_x^2}{s_x(ks_x)} = 1,$$

となり、予想通りの結果が得られます。

³厳密には、点が $x_i = \bar{x}$ または $y_i = \bar{y}$ を満たす場合、この比は使えません。その場合、「面積」は一次元の線に縮退します。

2.5.1 一つの点を動かすとどうなるか

r が幾何学的に何をしているのかをさらに理解するために、一つの点を少し上に動かし、データが直線上に並ばないようにします。



この操作によって、 r は1 から少し小さい値に減少すると予想されます。視覚的には、一つの点を直線から鉛直方向に動かすと、全体の平均 $\bar{y}$ が上に移動します。したがって、すべての距離 $y_i - \bar{y}$ も変化します。その結果、長方形は同じ縦横比を持たなくなります。上の図では、長方形の形が異なります。

代数的には、一つの点を鉛直方向に動かすことは、ある固定された点の y 成分に小さな値を加えることです。例えば、 $(x_4, y_4) \rightarrow (x_4, y_4 + \epsilon)$ とし、 ϵ を正の数とします。大まかに言えば、一つの点にこの変更を加えると共分散は少し増加します。しかし、標準偏差 s_y の公式には、 $(y_i - \bar{y})^2$ のように長さを二乗した項が含まれます。一つの点を ϵ だけ動かすと、この小さな変化は指数のために比較的大きな変化につながります。そのため、 s_y の項は $\text{cov}(x, y)$ の項より速く増加し、比 r は1 より小さくなります。

2.6 例： r の計算

第1節の通学時間のデータに戻ります。標本平均は

$$\bar{x} \approx 25.333 \quad \text{および} \quad \bar{y} \approx 34.333$$

です。

r を計算するためには、いくつかの種類の項を合計する必要があります。次の表に必要な情報を示します。

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
4	8	-21.333	-26.333	455.111	693.444	561.778
7	20	-18.333	-14.333	336.111	205.444	262.778
11	17	-14.333	-17.333	205.444	300.444	248.444
15	20	-10.333	-14.333	106.778	205.444	148.111
18	21	-7.333	-13.333	53.778	177.778	97.778
22	27	-3.333	-7.333	11.111	53.778	24.444
27	30	1.667	-4.333	2.778	18.778	-7.222
31	29	5.667	-5.333	32.111	28.444	-30.222
35	40	9.667	5.667	93.444	32.111	54.778
39	60	13.667	25.667	186.778	658.778	350.778
45	65	19.667	30.667	386.778	940.444	603.111
50	75	24.667	40.667	608.444	1653.778	1003.111
合計		0	0	2478.667	4968.667	3317.667

簡略化された計算公式を使うと、

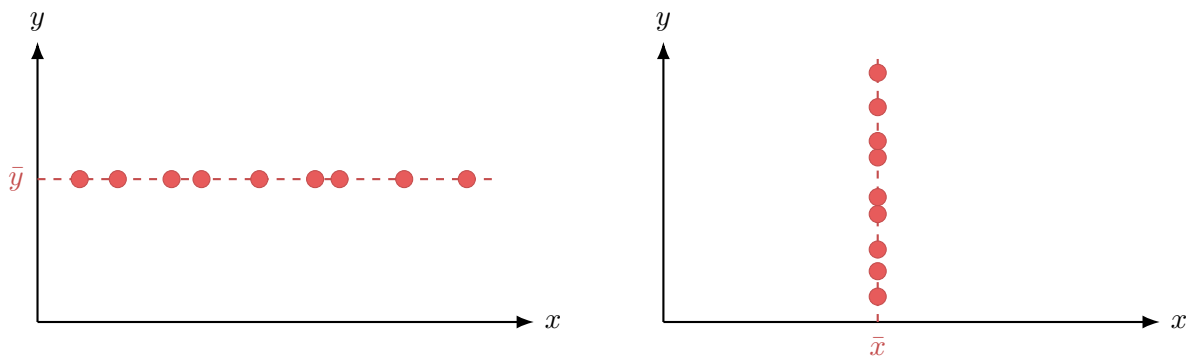
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{3317.667}{\sqrt{2478.667} \sqrt{4968.667}} \approx 0.945.$$

これは強い正の線形相関です。つまり、LUJ から遠くに住む学生ほど、通学時間が長くなる傾向があります。

3 相関の限界

3.1 r が定義されない場合

次の二つのデータ集合を考えます。



これらのデータ集合は、どちらも明らかに直線上にあります。しかし、標本相関係数 r は定義されません。どちらの場合も、一方の方向の広がりがゼロです。つまり、すべての y_i が $\bar{y}$ に等し

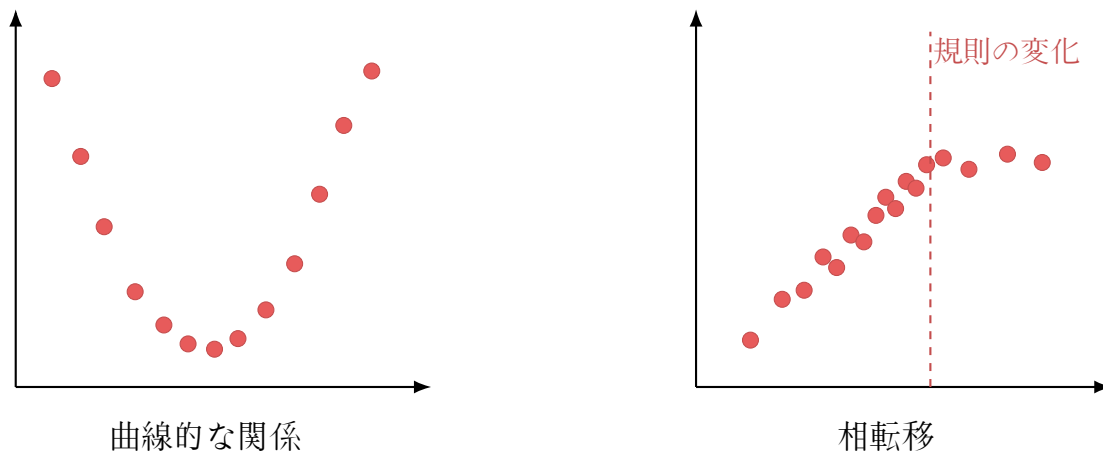
いか、すべての x_i が $\bar{x}$ に等しくなっています。そのため、共分散 $\text{cov}(x, y)$ と、 s_y または s_x の一方がゼロになり、

$$r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{0}{0},$$

となります。これは定義されません。

3.2 非線形相関

相関係数 r が測るのは線形相関だけです。特に、別の形で相関しているデータ集合であっても、 r の値が小さくなることがあります。また、ある範囲では関係が線形であっても、その後で別の種類の関係に変化する場合もあります。



科学的な応用では、ある関係が特定の点まではほぼ線形であっても、その後で新しい規則に変わることがあります。このような場合、相関係数は高い値を与える可能性がありますが、その値は系全体を正確に表しているとは限りません。

3.3 因果関係を伴わない相関

データに明らかな相関があるからといって、二つの変数の間に必ず因果関係があるとは限りません。例えば、 X と Y が相関している場合、次の可能性があります。

- X が Y の原因である。
- Y が X の原因である。
- X と Y が偶然によって相関している。
- X と Y の両方に影響する第三の変数が存在する。

ここで重要な標語は次のとおりです。

相関関係は因果関係を意味しない。

例として、アイスクリームの売上と日焼け止めの売上には相関が見られることがあります。これは、アイスクリームの売上が日焼け止めの売上を引き起こすという意味でも、その逆を意味するものでもありません。天候という第三の変数が、両方に影響しています。

相関関係は因果関係を意味しない

r の値は、線形パターンの存在、符号、強さを表すことができます。しかし、そのパターンに因果関係としての解釈を与えることはできません。