広島大学

データサイエンス基礎

Fundamental Data Science

第6回　相関と回帰

広島大学　数理統計学グループ

私たちがナビゲートします！

まさる
大 先生

今日の内容

- **2次元データにおける2つの変数の関係性を調べる手法**

  例 学生の身長と体重
  お店の来客数と売り上げ

- **要約統計量: 共分散, 相関係数**

  2つの変数の**直線関係**が数値でわかる!!

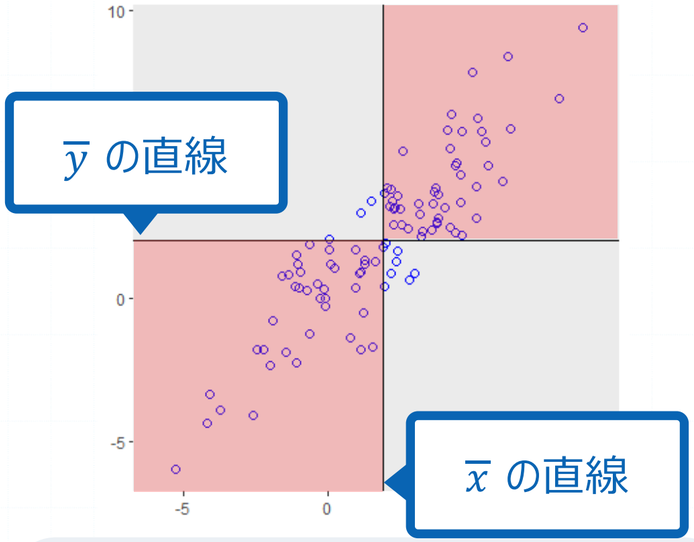
- **データ分析手法: 単回帰分析**

  片方の変数からもう一方の変数を**予測**できる!!

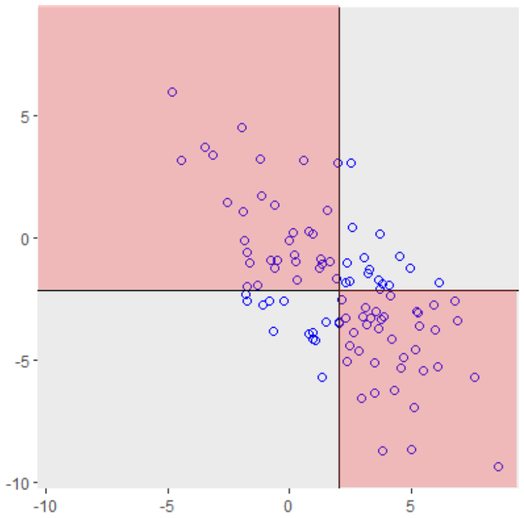
ひろこ
広子

共分散

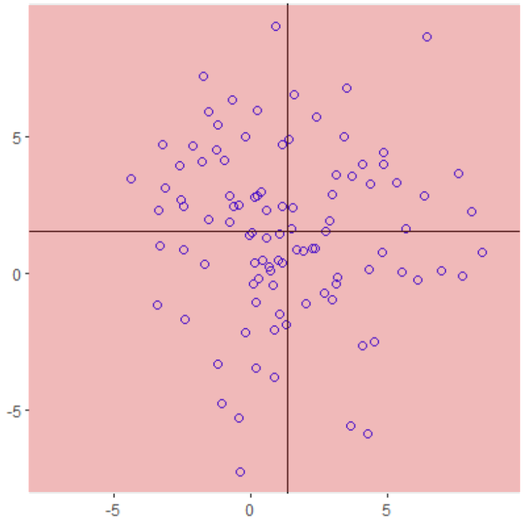
散布図 (復習)



右上と左下に
データ点が多い
⇒ 右上がりの (直線) 傾向



右下と左上に
データ点が多い
⇒ 右下がりの (直線) 傾向



4つの区画に
散らばっている
⇒ 直線的な関係はなさそう

散布図からわかる関係性を数値化したものが共分散

共分散

共分散 (covariance)

2つの変数 x, y の n 個のデータ $(x_1,y_1), \ldots, (x_n,y_n)$ に対する共分散:

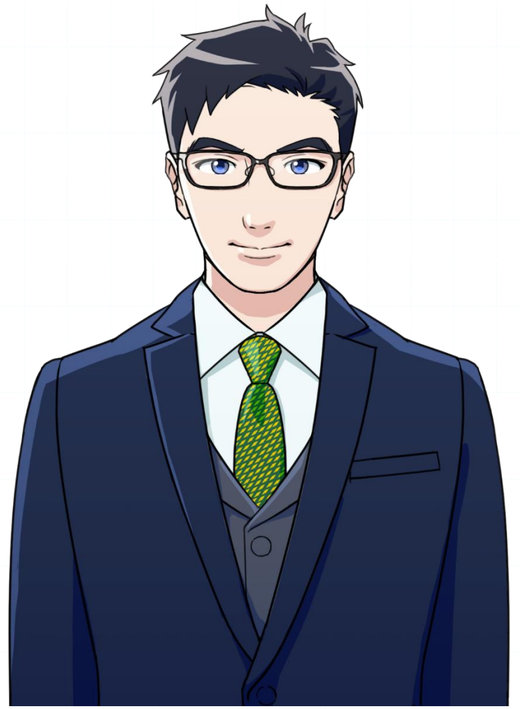
$$x, y \text{ の共分散 } s_{xy} = \frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{n}$$

$$\left(\bar{x} = \frac{x_1+\cdots+x_n}{n},\ \bar{y} = \frac{y_1+\cdots+y_n}{n} : \text{平均値} \right)$$

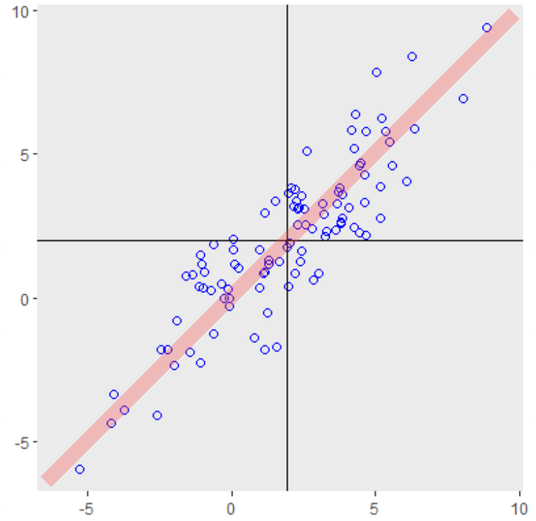
各変数から平均値を引いたときの掛け算の平均値

$$(x_i-\bar{x})(y_i-\bar{y})$$

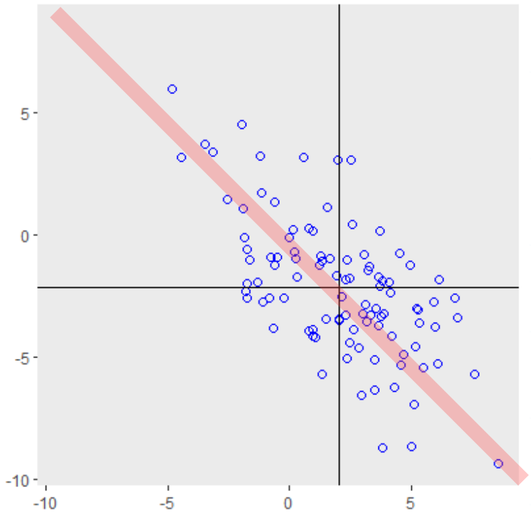
まずは計算式を
確認してみましょう



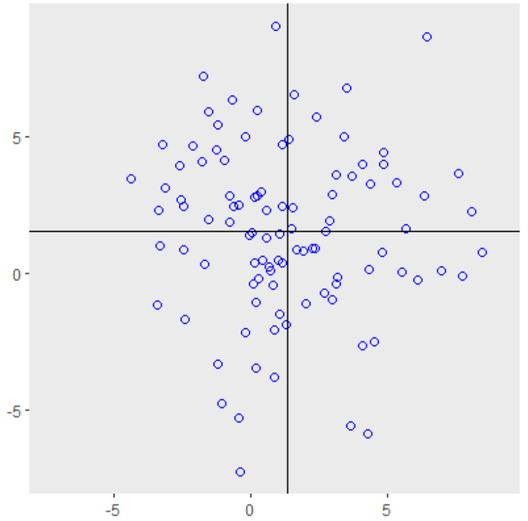
共分散の解釈



$s_{xy} = 6.53$



$s_{xy} = -5.02$



$s_{xy} = -0.35$

$s_{xy} > 0$ のとき右上がりの (直線) 傾向

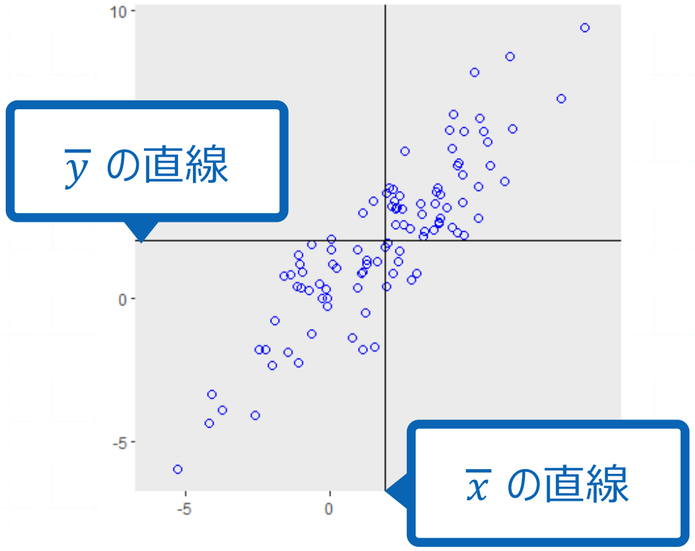
$s_{xy} < 0$ のとき右下がりの (直線) 傾向

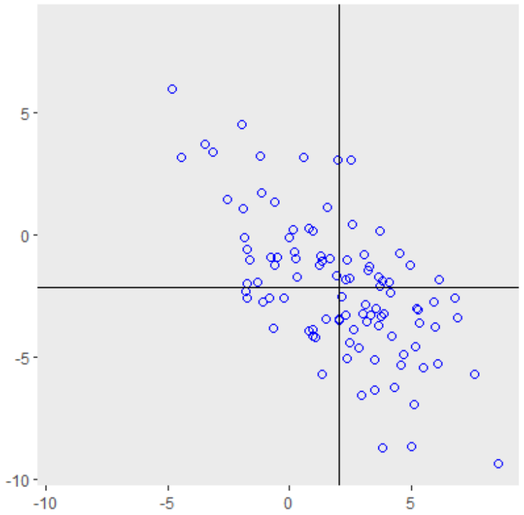
$s_{xy} = 0$ のとき直線関係はない

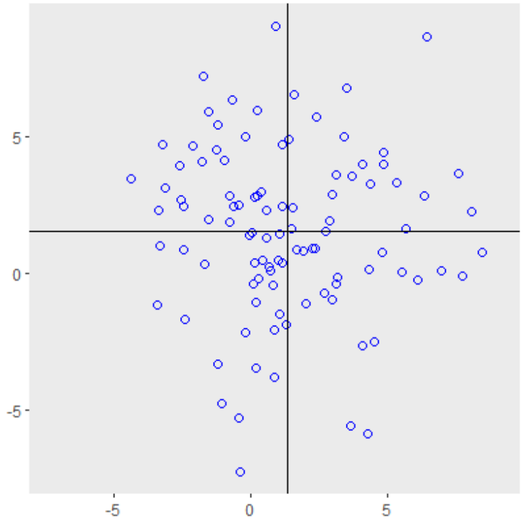
共分散を計算すれば,
直線の傾向がわかる

なぜこんな計算式??

$$s_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$







右上, 左下にデータが多い

$(x_i - \bar{x})(y_i - \bar{y}) > 0$
となるデータが多い

$s_{xy} > 0$ になりやすい

右下, 左上にデータが多い

$(x_i - \bar{x})(y_i - \bar{y}) < 0$
となるデータが多い

$s_{xy} < 0$ になりやすい

4つの区画に満遍なくデータが多い

$(x_i - \bar{x})(y_i - \bar{y})$が
正も負も同じくらいある

$s_{xy} = 0$ に近くなりやすい

相関係数

共分散の特徴

同じデータでも**単位を変える**と共分散の**値も変わる**

ID	身長 x (cm)	体重 y (**kg**)
1	152	44
2	160	49
3	165	54
4	168	59
5	170	61

$s_{xy} = 39.8$

$s_{xy} = 39800$

違う値!!

ID	身長 x (cm)	体重 y (**g**)
1	152	44000
2	160	49000
3	165	54000
4	168	59000
5	170	61000

単位を変えればいくらでも大きな (or 小さな) 値にできる

→ 共分散の値の大きさに意味はない!!

→ 直線関係がどれくらい強いかまではわからない

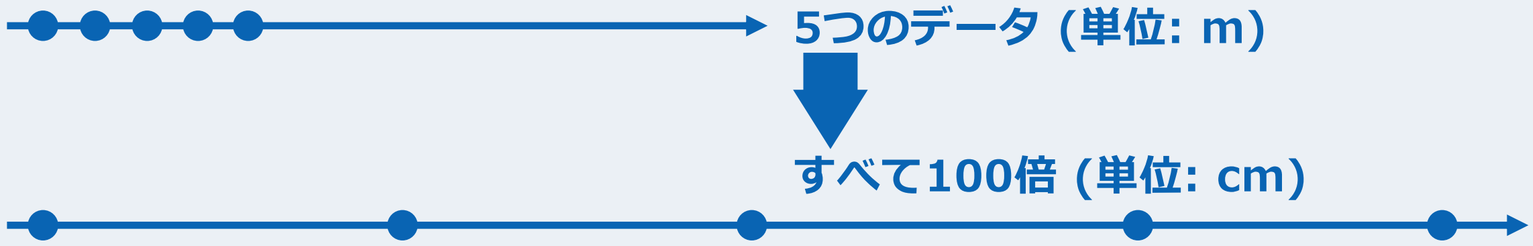
共分散の修正

単位によって値が変わらないよう共分散の計算を修正

$x_i - \bar{x}$ とか $y_i - \bar{y}$

平均値を引いたデータの
分散が単位を変えても**常に1となるよう**に変換

※ 単位の違いは分散の大きさに影響



データの間隔が拡がり分散は大きくなる

困りましたね…
計算方法を変えたら
上手くいきませんか?

標準化

平均値を引いたデータの
分散が単位を変えても**常に1となるよう**に変換

分散の平方根 (**標準偏差**) で割れば OK

$$\frac{x_i - \bar{x}}{s_x} \quad \frac{y_i - \bar{y}}{s_y} \quad (i = 1, \ldots, n)$$

これで単位を変えても
絶対に分散は1となる

- データから平均値を引く操作 $x_i - \bar{x}$ を**中心化**という
- データを中心化して標準偏差で割る操作 $\frac{x_i - \bar{x}}{s_x}$ を**標準化**という

相関係数

$$\frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{n}$$

$\frac{(x_1-\bar{x})}{s_x}\cdot\frac{(y_1-\bar{y})}{s_y}$ に置き換える!!

単位の影響を除いた共分散が相関係数

相関係数 (correlation coefficient)

2つの変数 x, y の n 個のデータ $(x_1,y_1),\ldots,(x_n,y_n)$ に対する相関係数:

x, y の相関係数 r_{xy}

$$= \left\{\frac{(x_1-\bar{x})}{s_x}\cdot\frac{(y_1-\bar{y})}{s_y}+\cdots+\frac{(x_n-\bar{x})}{s_x}\cdot\frac{(y_n-\bar{y})}{s_y}\right\}/n$$

標準化したデータの共分散

$$= \frac{s_{xy}}{s_x s_y}$$

(s_x: x の標準偏差, s_y: y の標準偏差)

※ $s_x = 0$ または $s_y = 0$ のときは, $r_{xy} = 0$ と定義する.

相関係数

単位が違っても相関係数は同じ値

ID	身長 x (cm)	体重 y (kg)
1	152	44
2	160	49
3	165	54
4	168	59
5	170	61

$s_{xy} = 39.8$

$r_{xy} = 0.98$

ID	身長 x (cm)	体重 y (g)
1	152	44000
2	160	49000
3	165	54000
4	168	59000
5	170	61000

$s_{xy} = 39800$

$r_{xy} = 0.98$

実は相関係数には,
共分散にはない
有用な性質があるんです!

相関係数の性質

性質1 $-1 \leq r_{xy} \leq 1$

相関係数は必ず**－1以上1以下**の値に収まる

性質2 $r_{xy} = 1$ ⇔ **データは傾きが正の直線上にある**

$r_{xy} = -1$ ⇔ **データは傾きが負の直線上にある**

$r_{xy} = 1, -1$ のときデータは**完全に直線上**

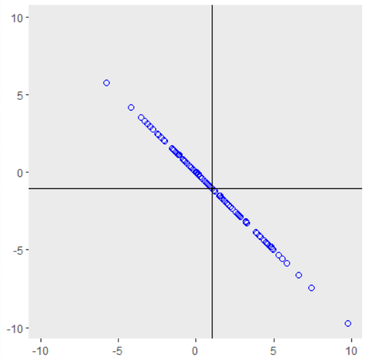
→ 相関係数で**直線関係の強さ**が測れる!!

- 1に近いほど**右上がりの直線関係**が強い
- −1に近いほど**右下がりの直線関係**が強い
- 0に近いほど直線関係は弱い

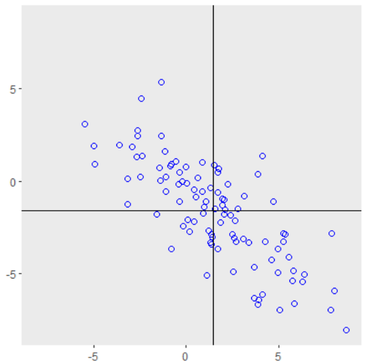
相関係数

特定の相関係数をもつデータの散布図

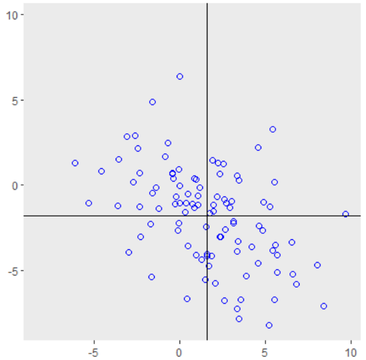
$r_{xy} = -1$



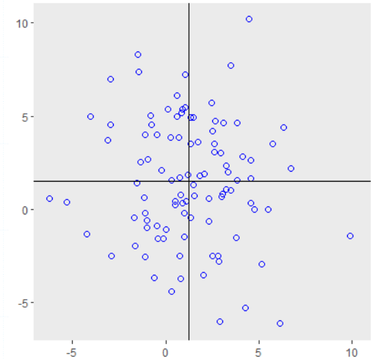
$r_{xy} = -0.8$



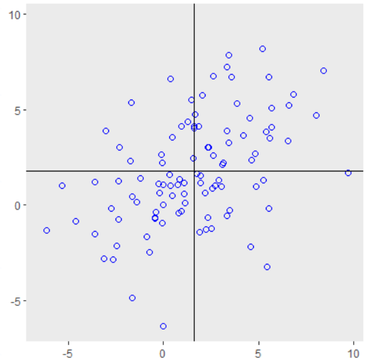
$r_{xy} = -0.5$



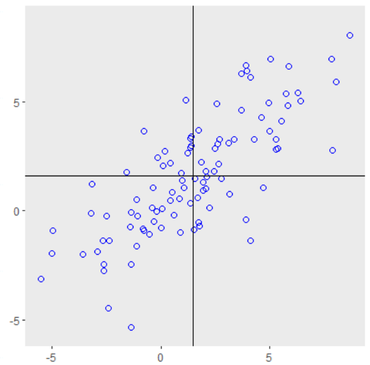
$r_{xy} = 0$



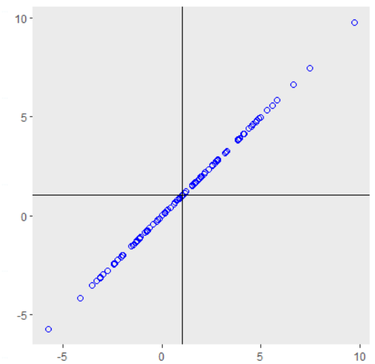
$r_{xy} = 0.5$



$r_{xy} = 0.8$



$r_{xy} = 1$

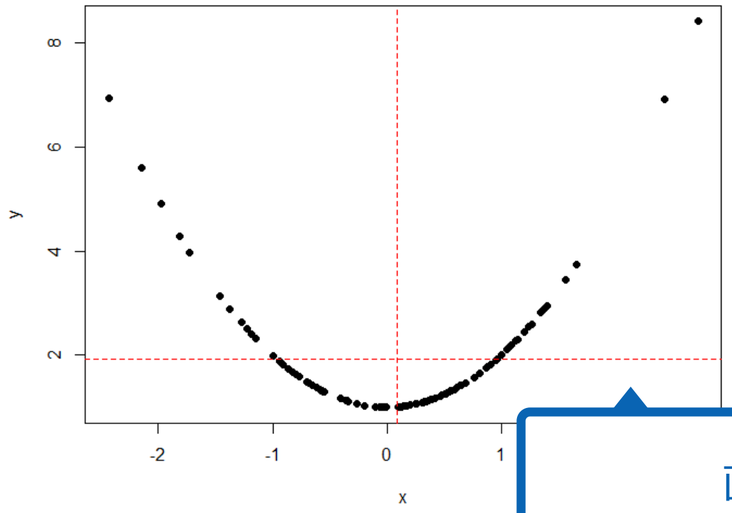


こうやって
直線関係の強さが測れる
わけですね!

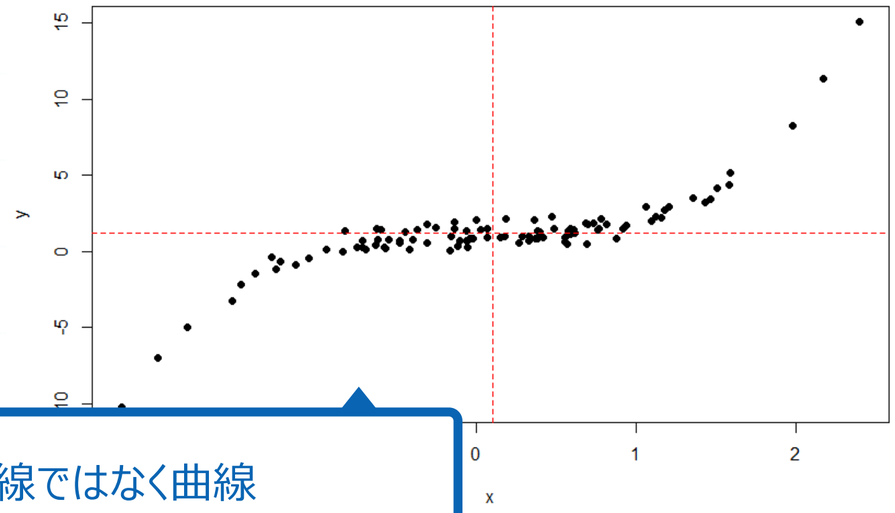
相関係数（注意）

相関係数の値だけで判断すると危険!!

$r_{xy} = 0$



$r_{xy} = 0.8$



直線ではなく曲線

相関係数と散布図の両方を確認すべき

直線以外の関係性があるかもしれないことに注意が必要です

相関係数の解釈

$r_{xy} > 0$ のとき**正の相関**（右上がりの直線傾向）

$r_{xy} < 0$ のとき**負の相関**（右下がりの直線傾向）

$r_{xy} = 0$ のとき**無相関**（直線関係はない）

- r_{xy} が1または−1に近いほど相関は強い
  中間は−0.7と0.7くらい
- **直線関係**の強さを表す

相関係数 r_{xy}	解釈
$-1 \leq r_{xy} \leq -0.7$	強い負の相関
$-0.7 \leq r_{xy} \leq -0.4$	やや負の相関
$-0.4 \leq r_{xy} \leq -0.2$	弱い負の相関
$-0.2 \leq r_{xy} \leq 0.2$	ほとんど相関なし
$0.2 \leq r_{xy} \leq 0.4$	弱い正の相関
$0.4 \leq r_{xy} \leq 0.7$	やや正の相関
$0.7 \leq r_{xy} \leq 1$	強い正の相関

回帰分析

モチベーション例

テストの点数を80点まで上げたい

勉強時間を増やせば点数は上がりそうだけど
80点とるにはどれくらい勉強すればいい??

でも勉強しすぎるのは嫌だ

Aさん

知り合い	勉強時間 (h)	点数
1	8	74
2	5.2	68
⋮	⋮	⋮
30	10	92

他の人に聞いた勉強に関する**データ**から
どれくらいの勉強時間で何点とれるか
知りたい

回帰分析

～時間で～点とれそう

回帰分析

回帰分析 (regression analysis)

入力 (**変数 x**) に対して**ある変換**を行い結果 (**変数 y**) を出力する方法

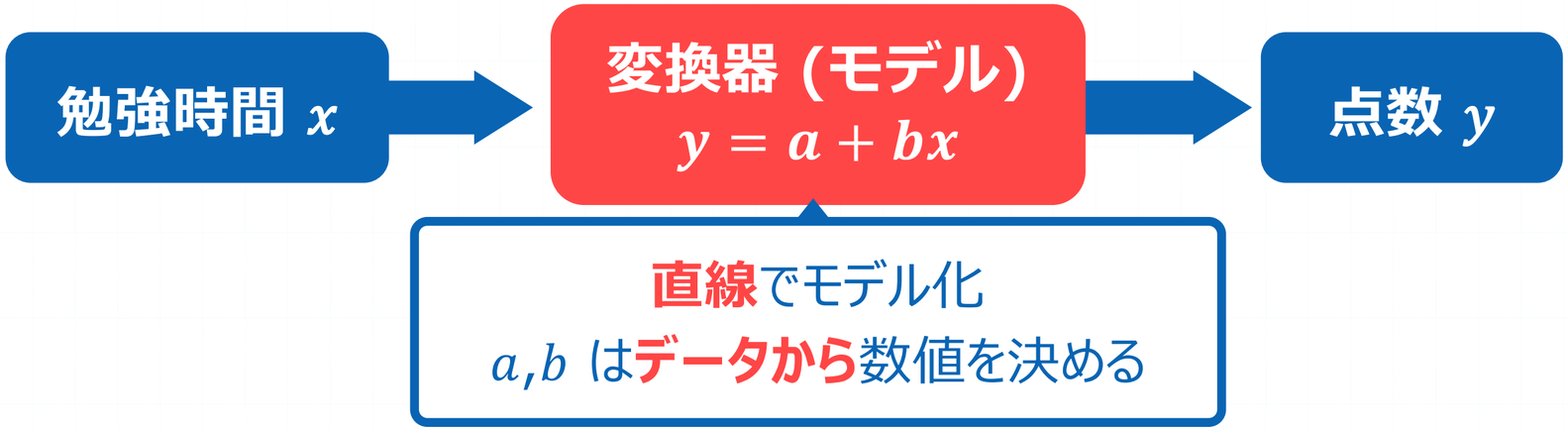
例

入力 (変数 x)	結果 (変数 y)
勉強時間	試験の点数
勉強時間	試験の点数, 部活時間
身長, 体重, 年齢	病気の疾患率

変換器（モデル）を
具体的に決めないと
分析はできない!!

回帰分析の例

知り合い	勉強時間 (h)	点数
1	8	74
2	5.2	68
⋮	⋮	⋮
30	10	92



$\hat{a}, \hat{b}$: データから決めた a, b の値

勉強時間が7時間なら，
点数は $\hat{a} + \hat{b} \times 7$点 とれるだろう

いろいろな回帰分析 (名前だけ紹介)

単回帰分析 → 拡張!!

- **多項式回帰分析**
- **重回帰分析**
- **多変量回帰分析**
- **ロジスティック回帰分析**
- **Lasso 回帰分析**
- **ディープラーニング**

なので，まずは
一番基本となる単回帰分析
について知りましょう

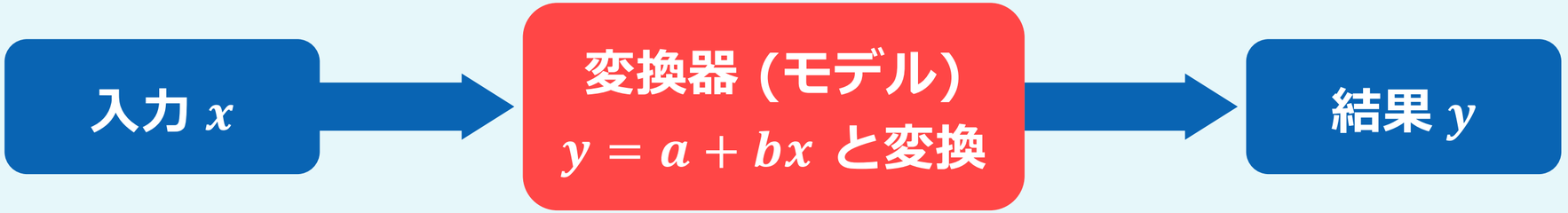
単回帰分析はいろんな回帰分析の基礎

単回帰分析
概要

単回帰分析

単回帰分析 (simple regression analysis)

1つの入力 x を**直線**によって変換し**1つ**の結果 y を出力する回帰分析



用語の定義

x を**説明変数** (予測変数, などともいう),

y を**目的変数** (結果変数, などともいう),

直線 $y = a + bx$ を**回帰直線**, a と b を**回帰係数**という

単回帰分析の目的

目的1

目的変数 y に対する
説明変数 x の影響を定量化する

回帰直線 $y = a + bx$ から
x が1増えれば y は b 増える

例
点数$= a + b \times$勉強時間
勉強時間が1時間増えれば点数が b 点増える

目的2

x に対する未知（将来）の y を予測する

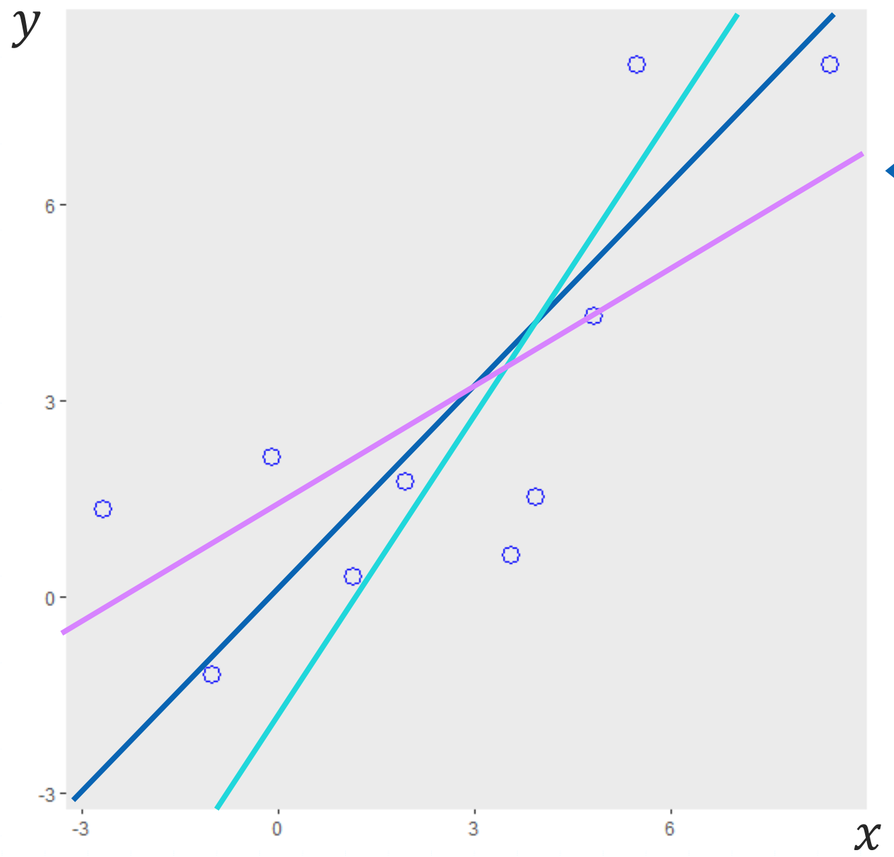
例
点数$= a + b \times$勉強時間
10時間勉強すれば, 点数は $(a + b \times 10)$ 点と予測できる

単回帰分析を行う目的は
大きく2つあります

単回帰分析
回帰係数の決め方

回帰係数 a,b の決め方

データ (x_i,y_i) $(i = 1,\ldots,n)$ から
回帰係数 a,b (つまり回帰直線) を決める



見た目だと…
ひとつに決めれない!!

何かしらの基準を定めて
ひとつの直線を求める!!

回帰直線って
どうやってかいたら
いいんでしょう…?

回帰係数 a,b の決め方

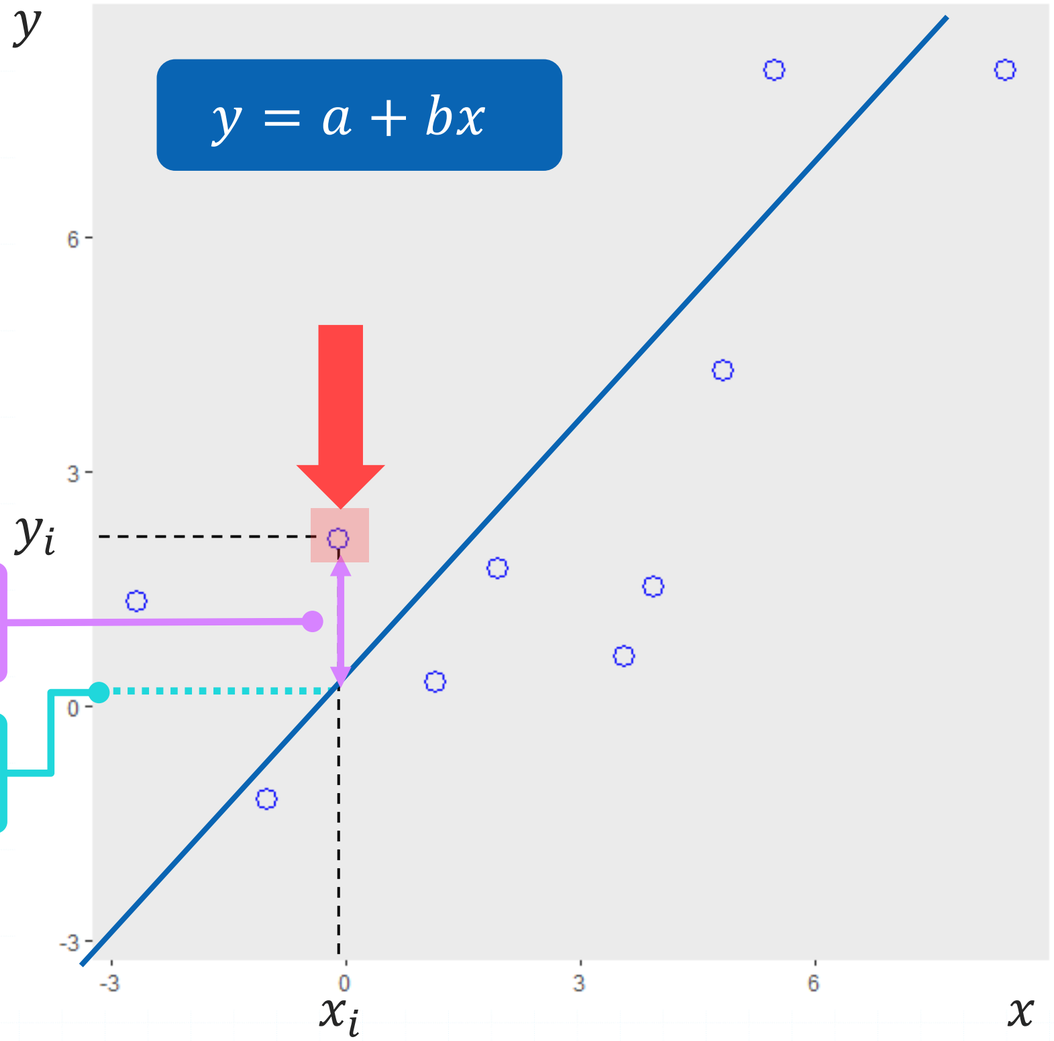
基準 各データに対する回帰直線の式 $a+bx_i$ と y_i が近いほどよい

➡ **残差** $y_i-(a+bx_i)$ が0に近いほどよい

全てのデータの残差について考える!!

$y_i-(a+bx_i)$

$a+bx_i$



最小2乗法

最小2乗法 (least squares method)

残差 $y_i - (a + bx_i)$ の2乗和 (**残差平方和**) が最小となるよう
回帰係数 a, b を決める方法

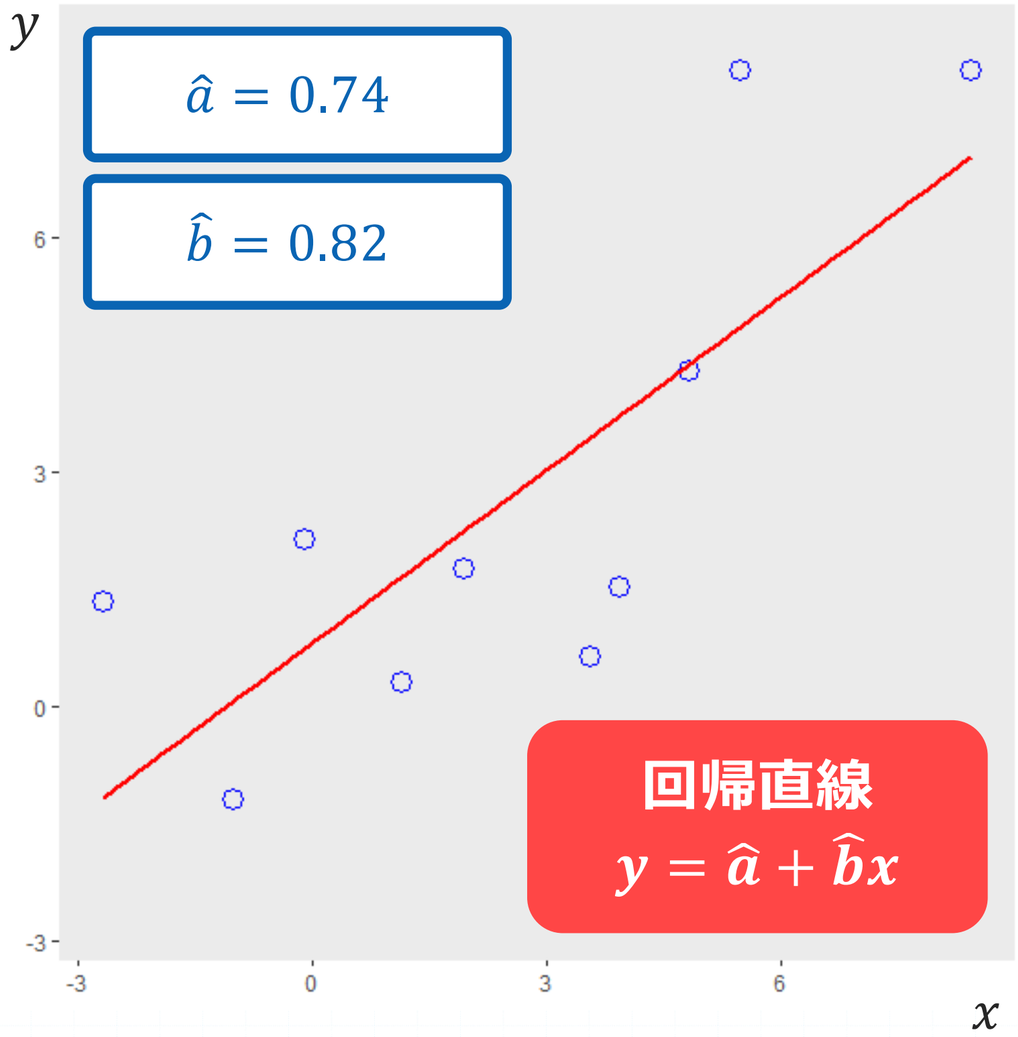
残差平方和: $\{y_1 - (a + bx_1)\}^2 + \cdots + \{y_n - (a + bx_n)\}^2$

残差平方和が
1番小さくなる a, b を求める!!

最小2乗推定値: $\hat{b} = r_{xy}\frac{s_y}{s_x}$, $\hat{a} = \bar{y} - \hat{b}\bar{x}$

$y = \hat{a} + \hat{b}x$ を回帰直線として用いる!!

回帰係数の当てはめ例



関係性の定量化

x が1増えれば
y は0.82増える

予測

x が10のとき
y は 0.74 + 0.82×10
= 8.94 と予測できる

回帰直線が引けたら,
単回帰分析の2つの目的も
実行することができます

練習問題 (1)

Q ある変数 (x,y) のデータに対して, 以下の値が得られたとする:

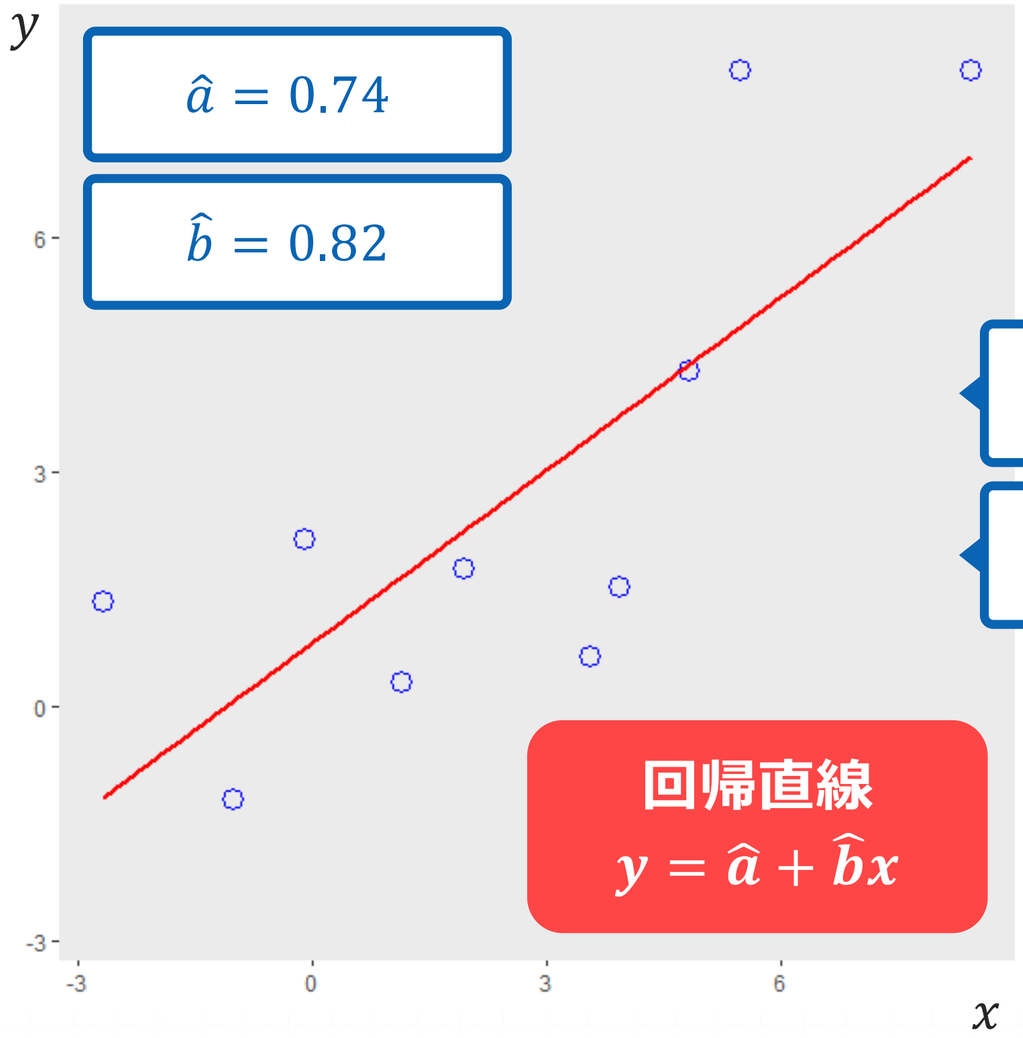
$$\bar{x} = 6, \quad \bar{y} = 6, \quad s_x^2 = \frac{26}{5}, \quad s_{xy} = \frac{14}{5}$$

y を目的変数, x を説明変数として単回帰分析を行うとき, 最小2乗法によって得られた回帰直線の式はどうなるか求めよ. (スライドp.12, 29を参考にせよ)

A

単回帰分析
決定係数

決定係数



求めた回帰直線 $y = \hat{a} + \hat{b}x$ が
データにどのくらい
当てはまっているかを知りたい

最小2乗法で求めたけど

実はあまり当てはまってないかも…??

当てはまり具合を
数値で知りたい!!

決定係数

決定係数 (coefficient of determination; R^2)

回帰直線 $y = \hat{a} + \hat{b}x$ の当てはまりの良さを表す指標

$$R^2 = 1 - \frac{(\hat{a} + \hat{b}x_1 - y_1)^2 + \cdots + (\hat{a} + \hat{b}x_n - y_n)^2}{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}$$

$$= 1 - \frac{(\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2}{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}$$

分子: **誤差変動**

分母: **全変動**

$\hat{y}_i = \hat{a} + \hat{b}x_i$: x_i に対する y の**予測値**

それでは
この決定係数は
どんな意味をもつでしょうか?

決定係数の性質と解釈

性質

$$0 \leq R^2 \leq 1$$

$$R^2 = 1 - \frac{(\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2}{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}$$

誤差変動

全変動

解釈

予測値が実際のデータに近いほど誤差変動は小さくなる

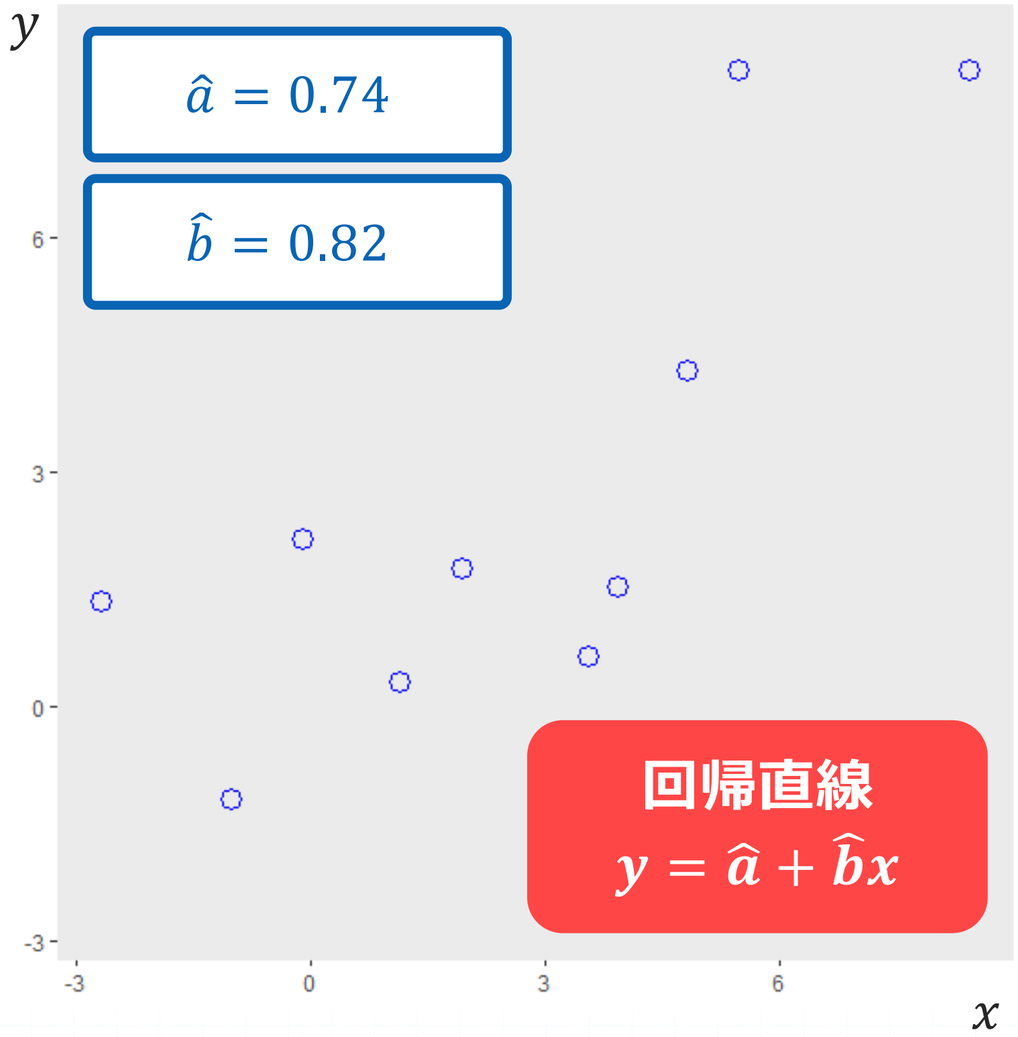
➡ R^2 は1になる

R^2 が1に近い ⇔ 当てはまりは良い

R^2 が0に近い ⇔ 当てはまりは悪い

中間は0.5

決定係数の使用例

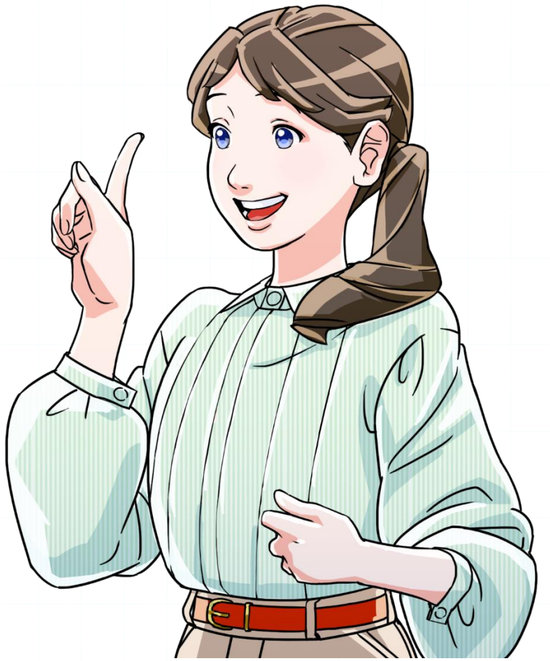


決定係数 $R^2 = 0.77$

当てはまりは良いから
予測結果も有効そう

**当てはまりの
良さを決定係数で
判断!!**

1に近いので
当てはまりはそこそこ
良さそうですね



単回帰分析に関する注意

決定係数が小さかった

単回帰分析は適していない可能性あり
直線ではなく曲線のほうが良い??
$\Longrightarrow y = a + b_1 x + b_2 x^2 \cdots + b_k x^k$ (**多項式回帰分析**)

複数の説明変数 $x_1, x_2, \ldots, x_k$ を用いて回帰分析??
$\Longrightarrow y = a + b_1 x_1 + b_2 x_2 \cdots + b_k x_k$ (**重回帰分析**)

※ 単回帰分析は**直線関係**のありそうなデータに対する
手法であることに注意

詳しくは
次回の講義で話します

見かけの相関 (spurious correlation) に注意

今日のまとめ

単回帰分析の手順例

❶ 説明変数 x と目的変数 y の散布図と相関係数から
直線関係がありそうかそうでないかを確認

❷ データから回帰係数 a, b を求める (例: 最小2乗法)

❸ 決定係数を計算して当てはまり具合を評価

❹ 予測などの分析を行う