Ｒで始めるデータサイエンス②

# Ｒを使って線形回帰

- 関係を数式で表す　～単回帰
- 単回帰の数理　～最小二乗法
- 単回帰の性質　～Ｒで単回帰
- 多項式に近似　～Yule-Walker
- 身近なデータに応用　～地球温暖化
- Excelデータ をRで処理



**課題１**

# 縦軸と横軸の関係は？

右表のデータについて
ｘとｙの相関係数

$r = \frac{s_{xy}}{s_x s_y}$　と決定係数

$r^2$　を求めよ

| | 横軸 | 縦軸 |
|---|---|---|
| i | x(i) | y(i) |
| 1 | 0 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 3 |

平均

$$\begin{cases} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \end{cases}$$

分散と共分散

$$\begin{cases} s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2 \\ s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{cases}$$

課題２

# 回帰直線は？

右表のデータについて
ｘからｙへの回帰直線

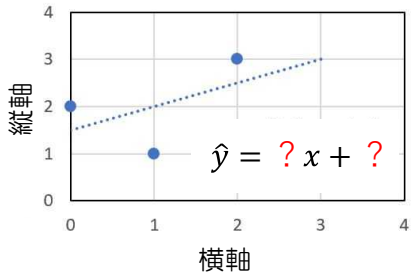
$\hat{y} = bx + a$　を求めよ

ただし、

$$b = \frac{s_{xy}}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$

| | 横軸 | 縦軸 |
|---|---|---|
| i | x(i) | y(i) |
| 1 | 0 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 3 |



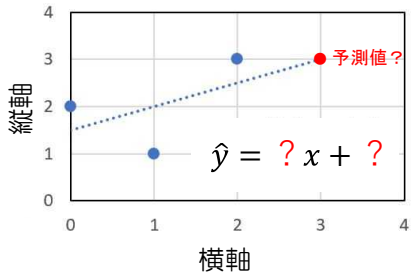



課題３

# 予測値は？

右表のデータについて
ｘからｙへの回帰直線
を使って、
xが３のときの
yの値を予測せよ

| | 横軸 | 縦軸 |
|---|---|---|
| i | x(i) | y(i) |
| 1 | 0 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 3 |
| 4 | 3 | **予測値？** |



**課題４**

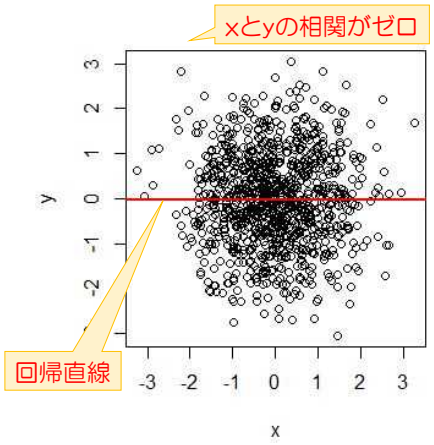
## 相関が無い時の回帰直線は？

x と y の相関が無い時の x から y への回帰直線を求めよ

$$\hat{y} = a + bx \qquad \begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{s_{xy}}{s_x^2} \end{cases}$$

において $x_i$ と $y_i$ の相関係数

$$r = \frac{s_{xy}}{s_x s_y}$$ をゼロとおく





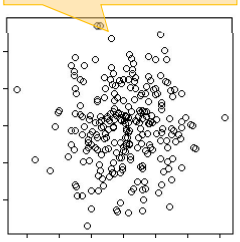

**Ｒで始めるデータサイエンス②**

# Ｒを使って線形回帰

- 関係を数式で表す　～単回帰
- 単回帰の数理　～最小二乗法
- 単回帰の性質　～Ｒで単回帰
- 多項式に近似　～Yule-Walker
- 身近なデータに応用　～地球温暖化
- Excelデータを**R**で処理

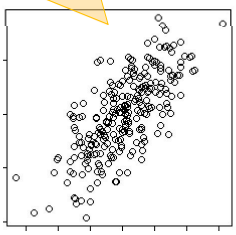
# 【前回】散布図をかいた

関係が無い



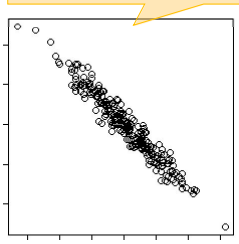
cor (x, y)
0.03
cor (x, y) ^2
0.00

関係がある



cor (x, x+y)
0.76
cor (x, x+y) ^2
0.58

負の関係が強い



cor (x, y−4*x)
−0.97
cor (x, y−4*x) ^2
0.94

相関係数

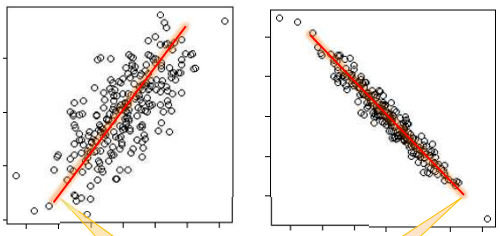
決定係数



# 【今回】直線をひきたい

学習事項

- Ｒで回帰直線をかく
- 回帰直線の切片と傾き
- 傾きと相関係数
- 予測誤差と決定係数



回帰直線

をＲでひきたい

# 理科と数学の関係は？

| 学生 | 理科 | 数学 |
|---|---|---|
| 1 | 85 | 78 |
| 2 | 96 | 90 |
| 3 | 64 | 62 |
| 4 | 90 | 82 |
| 5 | 76 | 70 |
| 6 | 80 | 75 |
| 7 | 73 | 70 |
| 8 | 70 | 63 |

**理科** Rに入力↓

r <- c(85,96,64,90,76,80,73,70)

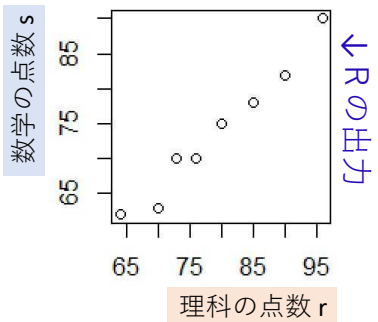
**数学** Rに入力↓

s <- c(78,90,62,82,70,75,70,63)



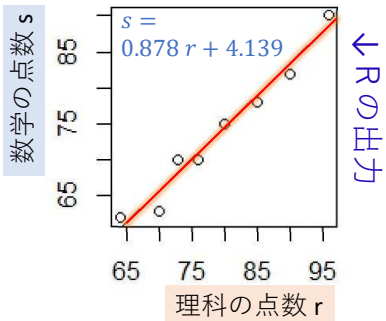
# Rで回帰直線をひく

plot(r,s) ←Rに入力




←Rの出力

m <- lm(s~r)
plot(r,s)
abline(m) ←Rに入力




←Rの出力

**参考**

```
> summary(m)
```

☜Rに入力

```
Call:
lm(formula = s ~ r)

Residuals:
     Min       1Q   Median       3Q      Max
-2.6251  -0.9696  -0.1047   1.5643   1.7398

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.13918    4.92647    0.84    0.433
r            0.87837    0.06168   14.24 7.49e-06 ***
```

目的変数 sを説明変数rでモデル化

残差（データと回帰直線の差）

回帰直線の係数

係数の有意性

$$s = 0.878\,r + 4.139$$

回帰直線



# 理科と英語の関係は？

| 学生 | 理科 | 英語 |
|---|---|---|
| 1 | 85 | 65 |
| 2 | 96 | 78 |
| 3 | 64 | 80 |
| 4 | 90 | 85 |
| 5 | 76 | 91 |
| 6 | 80 | 77 |
| 7 | 73 | 85 |
| 8 | 70 | 73 |

**理科** Rに入力↓

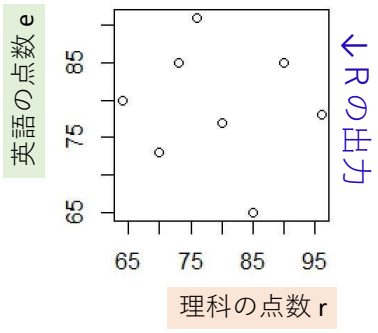
```
r <- c(85,96,64,90,76,80,73,70)
```

**英語** Rに入力↓

```
e <- c(65,78,80,85,91,77,85,73)
```

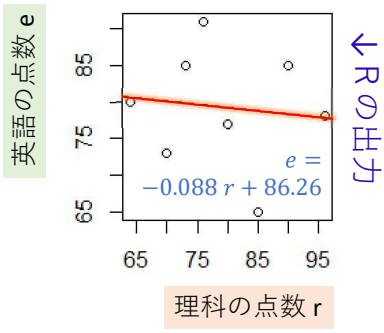
# Rで回帰直線をひく

plot(r,e) ←Rに入力




←Rの出力

m <- lm(e~r)
plot(r,e)
abline(m) ←Rに入力




←Rの出力



## 参考

```
> summary(m)
```

☜Rに入力

```
Call:
lm(formula = e ~ r)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7417  -3.4047  -0.1837   5.5732  11.4627

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.2558    24.4209   3.532   0.0123 *
r            -0.0884     0.3057  -0.289   0.7822
```

目的変数sを説明変数rでモデル化

残差（データと回帰直線の差）

回帰直線の係数

係数の有意性

$$e = -0.088\,r + 86.26$$

回帰直線

# 「あてはまり」の良さを測る

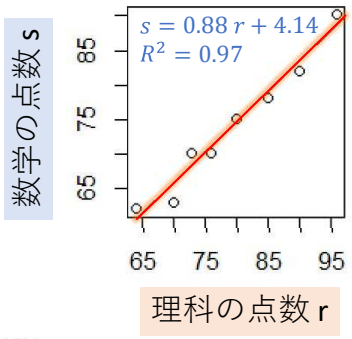
あてはまりが良い

> cor(r,s)

0.99 相関係数

> cor(r,s)^2

0.97 決定係数



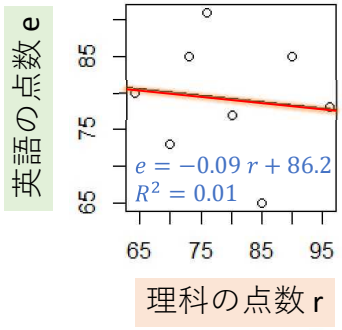

あてはまりが悪い

> cor(r,e)

-0.12 相関係数

> cor(r,e)^2

0.01 決定係数






# 回帰直線を R でかけた

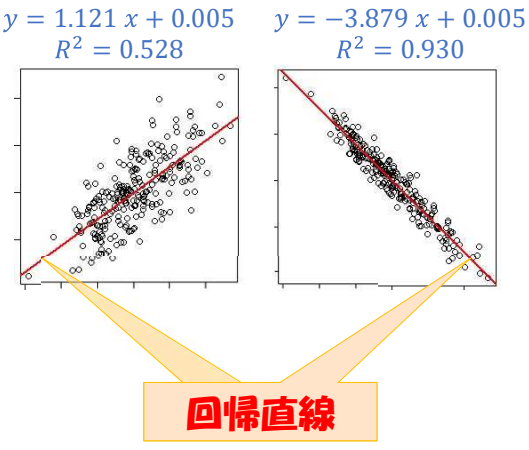
学習事項

✓ R で回帰直線をかく

・回帰直線の切片と傾き

・傾きと相関係数

・予測誤差と決定係数




**回帰直線**

を R でかけた

# Rのプログラム

これが無いと
毎回結果が異なる

```
set.seed(0)

x <- rnorm(256,0,1)
y <- rnorm(256,0,1)
w <- y-4*x
plot(x,w)

m <- lm(w~x)
abline(m,col="red",lwd=2)

summary(m)
```

正規分布する乱数
256点、平均0、分散1

横軸x、縦軸wの
散布図をかく

横軸xで縦軸wを
直線回帰する

回帰直線をかく
色は赤、太さは2

直線回帰の
統計量を示す

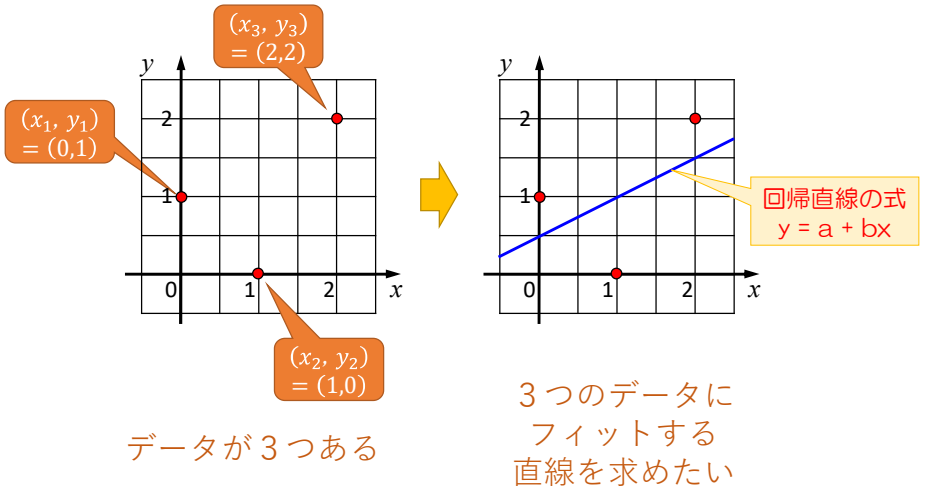


## Rで始めるデータサイエンス②

# Rを使って線形回帰

- 関係を数式で表す　～単回帰
- 単回帰の数理　～最小二乗法
- 単回帰の性質　～Rで単回帰
- 多項式に近似　～Yule-Walker
- 身近なデータに応用　～地球温暖化
- ExcelデータをRで処理

# 回帰直線の式を求めたい

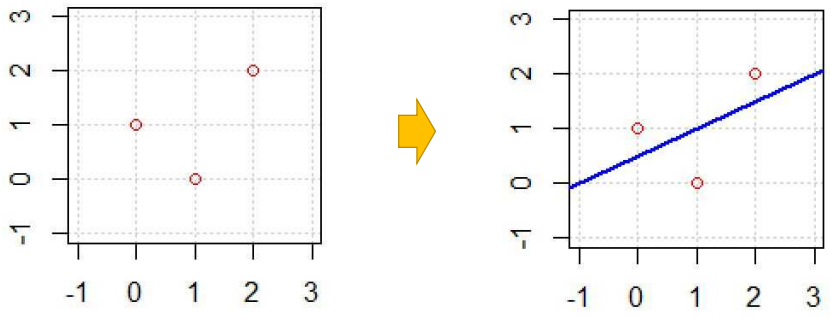



データが３つある

３つのデータに
フィットする
直線を求めたい

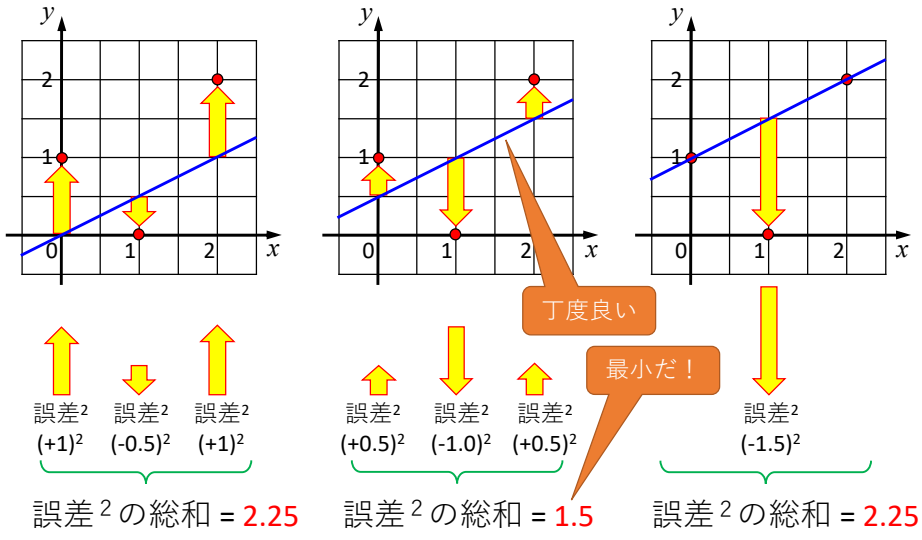


# Ｒのプログラム



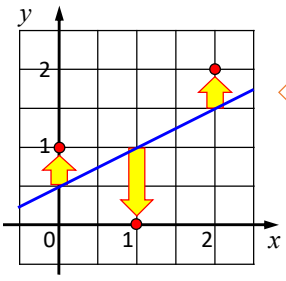
```
x=c(0,1,2)
y=c(1,0,2)
plot(x,y,xlim=c(-1,3),ylim=c(-1,3),panel.first=grid(NULL,NULL),col="red")
m <- lm(y~x)
abline(m,col="blue",lwd=2)
```

# 最小二乗法 の 考え方






# 最小二乗法 の 数理




回帰直線を $\hat{y} = a + bx$ とおく

$(x_1, y_1) = (0,1)$
$(x_2, y_2) = (1,0)$
$(x_3, y_3) = (2,2)$

データ（上図の赤丸）

誤差（黄色い矢印）の 2 乗の和

$$\sum_{i=1}^{3} \left(y_i - (a + bx_i)\right)^2$$

が最小となるように $a$ と $b$ を決める

**参考**

## 評価関数を $a$ で偏微分してゼロ

$$\frac{\partial I}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n}(a + b\cdot x_i - y_i)^2$$

$a$ で偏微分して 0 とおく

$$= \sum_{i=1}^{n} 2(a + b\cdot x_i - y_i)^1 \cdot \frac{\partial}{\partial a}(a + b\cdot x_i - y_i)$$

$$= \sum_{i=1}^{n} 2(a + b\cdot x_i - y_i)^1 \cdot \frac{\partial}{\partial a}(a)$$

$$= \sum_{i=1}^{n} 2(a + b\cdot x_i - y_i)^1 \cdot 1$$

$$= 0$$

零とおく

$$\sum_{i=1}^{n}(a + b\cdot x_i - y_i)^1 \cdot 1 = 0$$

$$I = \sum_{i=1}^{n}\{(a + bx_i) - y_i\}^2$$

評価関数



**参考**

## 評価関数を $b$ で偏微分してゼロ

$$\frac{\partial I}{\partial b} = \frac{\partial}{\partial b}\sum_{i=1}^{n}(a + b\cdot x_i - y_i)^2$$

$b$ で偏微分して 0 とおく

$$= \sum_{i=1}^{n} 2(a + b\cdot x_i - y_i)^1 \cdot \frac{\partial}{\partial b}(a + b\cdot x_i - y_i)$$

$$= \sum_{i=1}^{n} 2(a + b\cdot x_i - y_i)^1 \cdot \frac{\partial}{\partial b}(b\cdot x_i)$$

$$= \sum_{i=1}^{n} 2(a + b\cdot x_i - y_i)^1 \cdot x_i$$

$$= 0$$

零とおく

$$\sum_{i=1}^{n}(a + b\cdot x_i - y_i)^1 \cdot x_i = 0$$

$$I = \sum_{i=1}^{n}\{(a + bx_i) - y_i\}^2$$

評価関数

**参考**

# 連立方程式を解く （1/4）

$$\sum_{i=1}^{n}(a+b\cdot x_i-y_i)^1\cdot 1=0$$

$$\sum_{i=1}^{n}(a+b\cdot x_i-y_i)^1\cdot x_i=0$$

$$\Rightarrow \begin{aligned} a\sum_{i=1}^{n}1+b\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i&=0\\ a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_iy_i&=0\end{aligned}$$

回帰直線の
切片 a と傾き b
を求める

$$\begin{bmatrix}\sum_{i=1}^{n}1 & \sum_{i=1}^{n}x_n\\ \sum_{i=1}^{n}x_n & \sum_{i=1}^{n}x_i^2\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}y_i\\ \sum_{i=1}^{n}y_ix_i\end{bmatrix}$$



**参考**

# 連立方程式を解く （2/4）

$$\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}1 & \sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2\end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^{n}y_i\\ \sum_{i=1}^{n}x_iy_i\end{bmatrix}$$

$$=\frac{1}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\begin{bmatrix}\sum_{i=1}^{n}x_i^2 & -\sum_{i=1}^{n}x_i\\ -\sum_{i=1}^{n}x_i & n\end{bmatrix}\begin{bmatrix}\sum_{i=1}^{n}y_i\\ \sum_{i=1}^{n}x_iy_i\end{bmatrix}$$

$$=\frac{1}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\begin{bmatrix}\left(\sum_{i=1}^{n}x_i^2\right)\left(\sum_{i=1}^{n}y_i\right)-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_ix_i\right)\\ n\sum_{i=1}^{n}y_ix_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)\end{bmatrix}$$

**参考**

# 連立方程式を解く（3/4）

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i x_i\right) \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \end{bmatrix}$$

$$= \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)\bar{y} - \bar{x}\left(\frac{1}{n}\sum_{i=1}^{n} y_i x_i\right) \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{x}\bar{y} \end{bmatrix}$$

$$= \frac{1}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2} \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\right)\bar{y} - \bar{x}\left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right) \\ \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{bmatrix}$$



**参考**

# 連立方程式を解く（3/4）

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}\bar{x} \\ \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{bmatrix}$$

以上より、

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \bar{y} - b\bar{x} \\ \frac{s_{xy}}{s_x^2} \end{bmatrix}$$

回帰直線の切片 a と傾き b が求まった

但し、

$$\begin{cases} s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{cases}$$

Rで始めるデータサイエンス②

# Rを使って線形回帰

- 関係を数式で表す　～単回帰
- 単回帰の数理　～最小二乗法
- 単回帰の性質　～Rで単回帰
- 多項式に近似　～Yule-Walker
- 身近なデータに応用　～地球温暖化
- Excelデータを**R**で処理



# 回帰直線の計算法

$$\begin{cases} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \end{cases}$$

横軸 $x_i$ の平均

縦軸 $y_i$ の平均

$$\begin{cases} s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{cases}$$

横軸 $x_i$ の分散

横軸 $x_i$ と縦軸 $y_i$ の共分散

回帰直線 $\hat{y} = bx + a$

傾き $b = \frac{s_{xy}}{s_x^2}$

切片 $a = \bar{y} - b\bar{x}$

# 練習問題 1

右表のデータについて
回帰直線と予測値を求めよ

| | 横軸 | 縦軸 |
|---|---|---|
| i | x(i) | y(i) |
| 1 | 0 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 3 |
| 4 | 3 | **予測値？** |

ただし、回帰直線を

$$\hat{y} = bx + a$$

とすると、

$$b = \frac{s_{xy}}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$

平均

$$\begin{cases} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \end{cases}$$

分散と共分散

$$\begin{cases} s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{cases}$$



# 練習問題 1 （解説1/3）

$$\begin{cases} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0+1+2}{3} = 1 \\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{2+1+3}{3} = 2 \end{cases}$$

| | 横軸 | 縦軸 |
|---|---|---|
| i | x(i) | y(i) |
| 1 | 0 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 3 |
| 4 | 3 | |

$$\begin{cases} s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{(0-1)^2 + (1-1)^2 + (2-1)^2}{3} = \frac{2}{3} \\ s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{(0-1)(2-2) + (1-1)(1-2) + (2-1)(3-2)}{3} = \frac{1}{3} \end{cases}$$

$$b = \frac{s_{xy}}{s_x^2} = \frac{1/3}{2/3} = 0.5$$

$$a = \bar{y} - b\bar{x} = 2 - \frac{1}{2} \cdot 1 = 1.5$$

回帰直線は？

# 練習問題１（解説2/3）

$$b = \frac{s_{xy}}{s_x^2} = 0.5$$

$$a = \bar{y} - \hat{b}\bar{x} = 1.5$$

以上より、回帰直線は

$$\hat{y} = bx + a$$

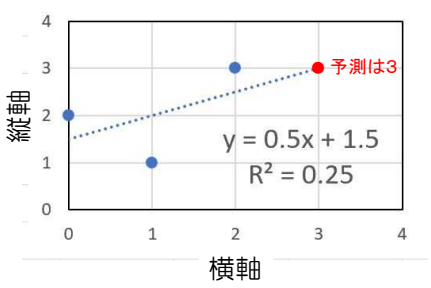
$$= 0.5\,x + 1.5$$

このとき、$x$=3 ならば

$$y = 0.5 \cdot 3 + 1.5$$

$$= 3$$ ← 予測値

| | 横軸 | 縦軸 |
|---|---|---|
| i | x(i) | y(i) |
| 1 | 0 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 3 |
| 4 | 3 | 予測は３ |






# 練習問題１（解説3/3）

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{2+1+3}{3} = 2$$

| | 横軸 | 縦軸 |
|---|---|---|
| i | x(i) | y(i) |
| 1 | 0 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 3 |
| 4 | 3 | |

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{(2-2)^2 + (1-2)^2 + (3-2)^2}{3} = \frac{2}{3}$$

$= s_x^2$ となっていた

$$r = \frac{s_{xy}}{s_x s_y} = \frac{1/3}{\sqrt{2/3}\sqrt{2/3}} = 0.5$$ ← 相関係数　　決定係数→ $r^2 = 0.25$

$$b = \frac{s_{xy}}{s_x^2} = \frac{1/3}{2/3} = 0.5$$ ← 回帰直線の傾き

## 縦軸と横軸に相関が無い場合は
# 回帰直線 ＝ 縦軸の平均

回帰直線は、

$$\hat{y} = a + bx \qquad \begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{s_{xy}}{s_x^2} \end{cases}$$

$$\begin{cases} \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i \end{cases} \qquad \begin{cases} s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s_{xy} = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{cases}$$

$x_i$ と $y_i$ の相関係数は

$$r = \frac{s_{xy}}{s_x s_y} \quad \Rightarrow \quad b = \frac{s_y}{s_x} r$$

$x_i$ と $y_i$ が無相関（$r = 0$）のとき

$$b = 0 \quad \Rightarrow \quad \hat{y} = \bar{y}$$

R で検証しよう

$y_i$ と $x_i$ に相関が無い場合は
回帰直線 ＝ $y_i$ の平均

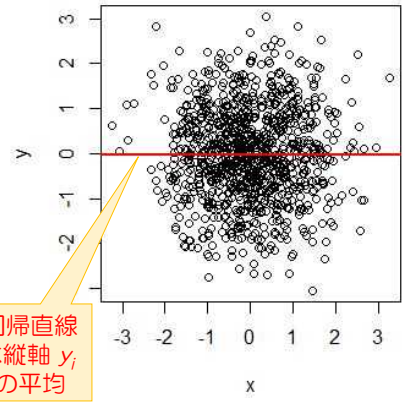


# R で検証しよう

```
set.seed(0)

x <- rnorm(1024,0,1)
y <- rnorm(1024,0,1)
plot(x,y)

m <- lm(y~x)
abline(m,col="red",lwd=2)

summary(m)
```

cor(x,y)
-0.00185

相関係数
ほぼゼロ




回帰直線
は縦軸 $y_i$
の平均

R に入力

R の出力

参考

## 縦軸の分散と横軸の分散が同じ場合は
# 回帰直線の傾き ＝ 相関係数

回帰直線は、

$$\hat{y} = a + bx \qquad \begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{s_{xy}}{s_x^2} \end{cases}$$

$$\begin{cases} \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i \end{cases}$$

$$\begin{cases} s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s_{xy} = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{cases}$$

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

$x_i$と$y_i$の相関係数は

$$r = \frac{s_{xy}}{s_x s_y} \quad \Rightarrow \quad b = \frac{s_y}{s_x} r$$

$x_i$の分散と$y_i$の分散が同じならば

$$s_x = s_y \quad \Rightarrow \quad b = r$$

$y_i$ の分散と $x_i$ の分散が同じ場合は
回帰直線の傾き ＝ 相関係数



# R で検証しよう

赤文字：Rに入力
青文字：Rの出力

```
set.seed(0)

x <- rnorm(1024,0,1)
y <- rnorm(1024,0,1)
w1 <- x+y
w2 <- w1/sd(w1)*sd(x)
```

```
var(x)
1.001069
var(w1)
2.066758
var(w2)
1.001069
```

xの分散とw1の分散は**異なる**

```
cor(w1,x)
0.6946333

m <- lm(w1~x)
coef(m)
(Intercept)         x
-0.01448707  0.99808708
```

相関係数 ↑ 異なる ↓ 傾き

xの分散とw2の分散は**同じ**

```
cor(w2,x)
0.6946333

m <- lm(w2~x)
coef(m)
(Intercept)         x
-0.01008249  0.69463335
```

相関係数 ↑ 同じ ↓ 傾き

参考

## 当てはまりが良い ⇔ 予測が当たる
## 決定係数が 1 ⇔ 予測誤差が 0

横軸 $x_i$ に対する縦軸の予測値

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

予測誤差の定義

決定係数の定義

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

予測誤差が 0 ならば、決定係数は 1
（予測が当たると）

※逆も然り



## R で検証しよう

```
set.seed(0)
x <- rnorm(1024,0,1)
y <- rnorm(1024,0,1)
```

```
w1 <- x*4+y
w2 <- x*4+y/100
```

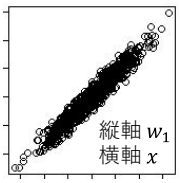
$$w_1 = 4x + y$$
$$w_2 = 4x + y/100$$

```
m <- lm(w1~x)
summary(m)
```

予測誤差が小さい

```
Residual standard error: 1.035 on 1022 degrees of freedom
Multiple R-squared:  0.9374,
F-statistic: 1.529e+04 on 1 and 1022 DF
```

決定係数が 1 に近い

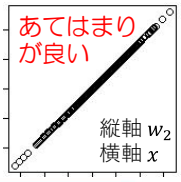



```
m <- lm(w2~x)
summary(m)
```

予測誤差がほぼ 0

```
Residual standard error: 0.01035 on 1022 degrees of freedom
Multiple R-squared:  1,
F-statistic: 1.531e+08 on 1 and 1022 DF
```

決定係数が 1



参考

## 最小二乗法による直線回帰の場合は
# 決定係数 ＝ 相関係数の２乗

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

決定係数

$$\hat{y}_i = \hat{a} + \hat{b}x_i \qquad \bar{y} = \hat{a} + \hat{b}\bar{x} \qquad \hat{b} = \frac{s_{xy}}{s_x^2} \qquad r = \frac{s_{xy}}{s_x s_y}$$

相関係数

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}\left(\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\bar{x}\right)^2}{ns_y^2} \\ &= \frac{\hat{b}^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{ns_y^2} \\ &= \frac{\hat{b}^2 s_x^2}{s_y^2} = \left(\frac{s_{xy}}{s_x^2}\right)^2 \frac{s_x^2}{s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} \end{aligned}$$

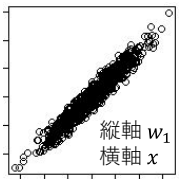
$$\therefore \quad R^2 = r^2$$

*R* は重相関係数



# R で検証しよう

```
set.seed(0)
x <- rnorm(1024,0,1)
y <- rnorm(1024,0,1)
w1 <- x*4+y
plot(x,w1)
```




```
m <- lm(w1~x)
summary(m)

Call:
lm(formula = w1 ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-3.02811 -0.70166 -0.00197  0.74541  3.05423

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01449    0.03235  -0.448    0.654
x            3.99809    0.03233 123.656   <2e-16 ***
---
Residual standard error: 1.035 on 1022 degrees of freedom
Multiple R-squared:  0.9374,    Adjusted R-squared:  0.9373
F-statistic: 1.529e+04 on 1 and 1022 DF,  p-value: < 2.2e-16

cor(x,w1)^2
0.9373501
```

相関係数の2乗 ←同じ→ 決定係数

Ｒで始めるデータサイエンス②

# Ｒを使って線形回帰

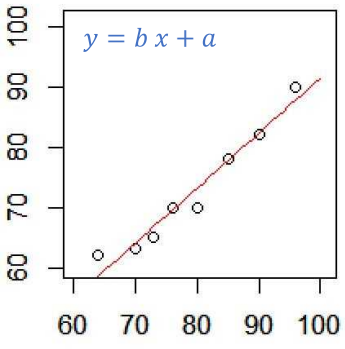
- 関係を数式で表す　～単回帰
- 単回帰の数理　～最小二乗法
- 単回帰の性質　～Ｒで単回帰
- 多項式に近似　～Yule-Walker
- 身近なデータに応用　～地球温暖化
- Excelデータをｒで処理



# １次の多項式に回帰

```
x <- c(85,96,64,90,76,80,73,70)
y <- c(78,90,62,82,70,70,65,63)
plot(x,y,xlim=c(60,101),ylim=c(60,101))

f    <- y ~ b*x +a
obj <- nls(f,start=c(b=0,a=0))

c  <- coefficients(obj)
xx <- seq(60,100,by=1)
yy <- c[1]*xx +c[2]

par(new=T)
plot(xx,yy,col="red",type="l",xlim=c(60,101),ylim=c(60,101))
```

*b, a*



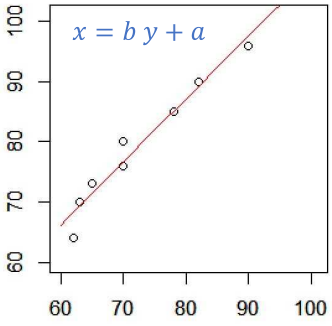

```
y ~ b * x + a
b=0. 9129
a=0. 1564
```

```
       b        a
0. 9129   0. 1564
```

# 1次の多項式（縦横入れ替え）

```
y <- c(85,96,64,90,76,80,73,70)
x <- c(78,90,62,82,70,70,65,63)
plot(x,y,xlim=c(60,101),ylim=c(60,101))

f    <- y ~ b*x +a
obj <- nls(f,start=c(b=0,a=0))

c  <- coefficients(obj)
xx <- seq(60,100,by=1)
yy <- c[1]*xx +c[2]

par(new=T)
plot(xx,yy,col="red",type="l",xlim=c(60,101),ylim=c(60,101))
```

*b, a*



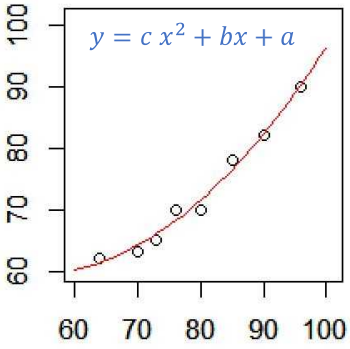

```
y ~ b * x + a
b=1.045977
a=3.416667

      b        a
1.045977 3.416667
```



# 2次の多項式に回帰

```
x <- c(85,96,64,90,76,80,73,70)
y <- c(78,90,62,82,70,70,65,63)
plot(x,y,xlim=c(60,101),ylim=c(60,101))

f    <- y ~ c*x^2 +b*x +a
obj <- nls(f,start=c(c=0,b=0,a=0))

c  <- coefficients(obj)
xx <- seq(60,100,by=1)
yy <- c[1]*xx^2+c[2]*xx+c[3]

par(new=T)
plot(xx,yy,col="red",type="l",xlim=c(60,101),ylim=c(60,101))
```

*c, b, a*

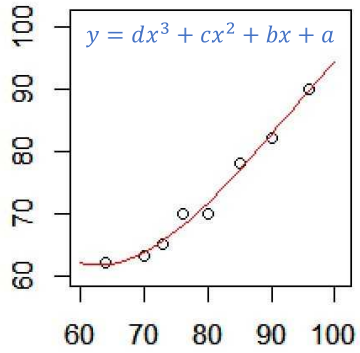



```
          c           b           a
0.01666556 -1.76031421 105.67443819
```

# 3次の多項式に回帰

```
x <- c(85,96,64,90,76,80,73,70)
y <- c(78,90,62,82,70,70,65,63)
plot(x,y,xlim=c(60,101),ylim=c(60,101))

f    <- y ~ d*x^3 +c*x^2 +b*x +a
obj <- nls(f,start=c(d=0,c=0,b=0,a=0))

c  <- coefficients(obj)
xx <- seq(60,100,by=1)
yy <- c[1]*xx^3+c[2]*xx^2+c[3]*xx+c[4]

par(new=T)
plot(xx,yy,col="red",type="l",xlim=c(60,101),ylim=c(60,101))
```

**d, c, b, a**




```
d                c                b                a
-4.777351e-04    1.310101e-01    -1.079047e+01    3.409658e+02
```



# 多項式（m-1次）に回帰

$$y = a_0 + a_1 x + a_2 x^2 \cdots + a_{m-1} x^{m-1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^{m-1} \\ 1 & x_2 & \cdots & x_2^{m-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^{m-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{m-1} \end{bmatrix}$$

$(\boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{a})$ の２乗の総和を最小にする

$$\boldsymbol{a} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

Yule-Walker方程式

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{m-1} \end{bmatrix}$$

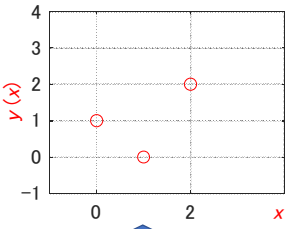
$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^{m-1} \\ 1 & x_2 & \cdots & x_2^{m-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^{m-1} \end{bmatrix}$$

n行×m列
**縦長** (n>m)

変数が　m 個
方程式が n 個

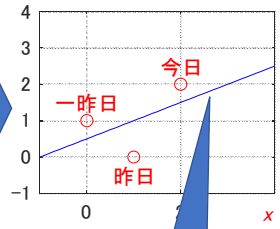
# 線形回帰で未来予測

**ステップ1**
データの散布図をかく




| $x$ | $y(x)$ |
|---|---|
| 0 | 1 |
| 1 | 0 |
| 2 | 2 |
| 3 | ? |

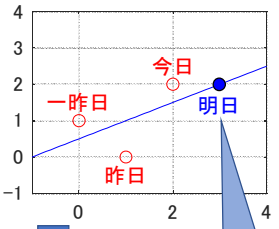
**ステップ2**
回帰直線をひく




これが
回帰直線

$$y(x) = 0.5 + 0.5\,x$$

**ステップ3**
未来を予測する



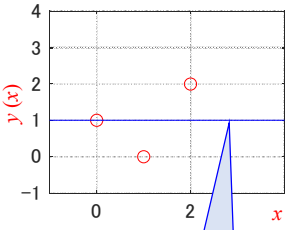

| $x$ | $y(x)$ |
|---|---|
| 0 | 1 |
| 1 | 0 |
| 2 | 2 |
| 3 | 2 |

未来を
予測できた



# 直線から曲線へ

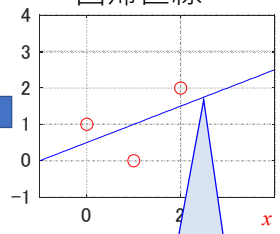
0 次の
回帰直線




$y(x) = 1.0$

0 次
パラメータが 1 個
データが 3 個

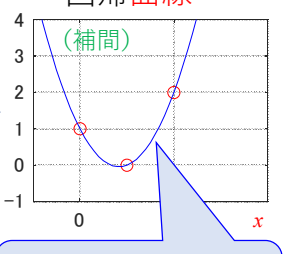
1 次の
回帰直線




$y(x) = 0.5 + 0.5\,x$

1 次
パラメータが 2 個
データが 3 個

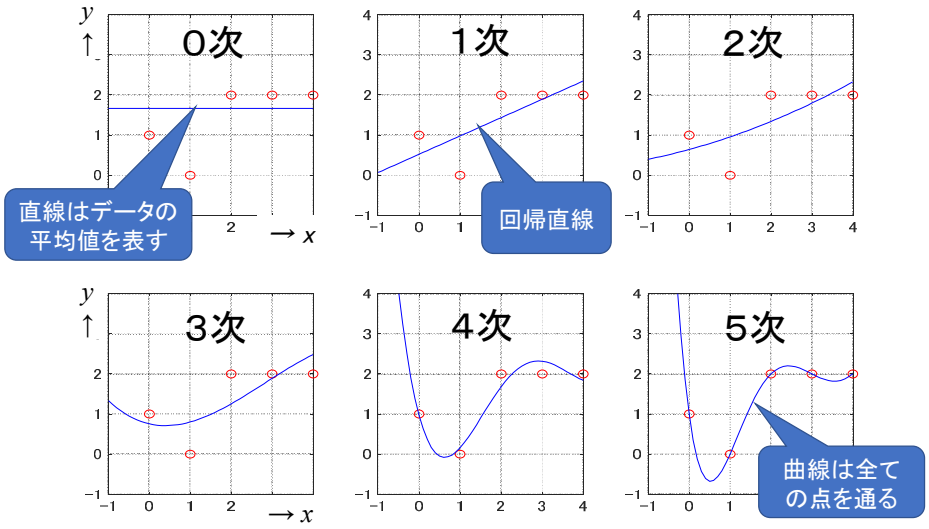
2 次の
回帰曲線




$y(x) = 1.0 - 2.5\,x + 1.5\,x^2$

2 次
パラメータが 3 個
データが 3 個

# 次数を増やすと…



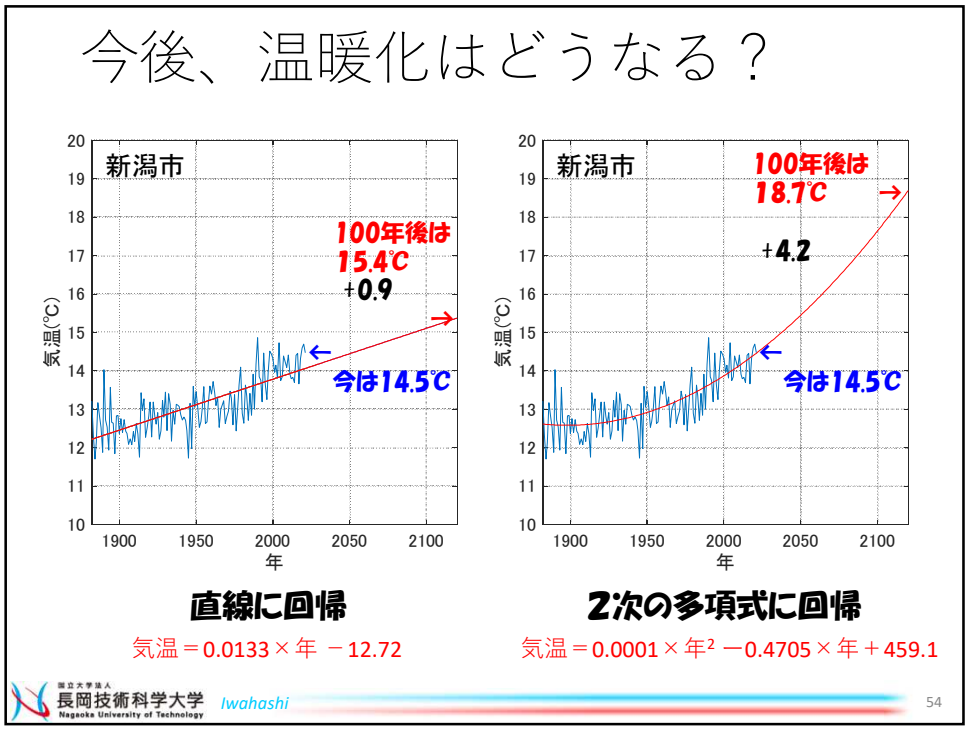



## Rで始めるデータサイエンス②

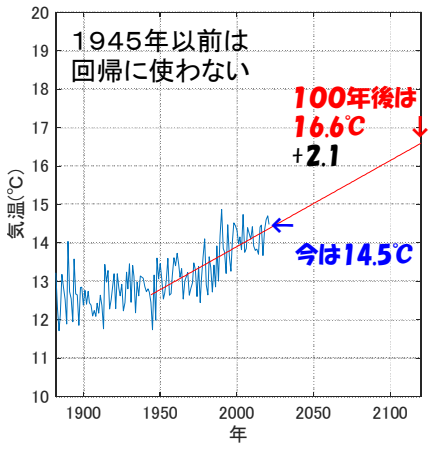
# Rを使って
# 線形回帰

- 関係を数式で表す　～単回帰
- 単回帰の数理　～最小二乗法
- 単回帰の性質　～Rで単回帰
- 多項式に近似　～Yule-Walker
- 身近なデータに応用　～地球温暖化
- Excelデータを R で処理

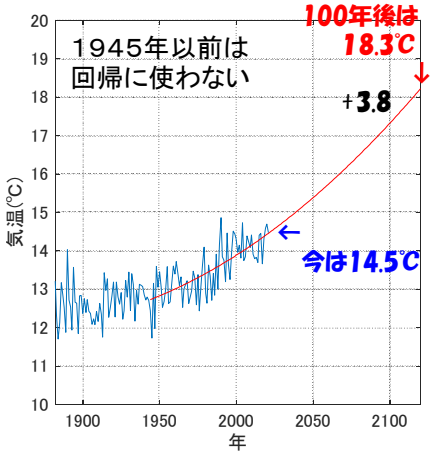

国土交通省
気象庁
Japan Meteorological Agency
https://www.data.jma.go.jp/gmd/risk/obsdl/index.php
ホーム
防災情報
各種データ・資料
地域の情報
ホーム > 各種データ・資料 > 過去の気象データ・ダウンロード
過去の気象データ・ダウンロード
重要なお知らせ
このページの使い方
検索条件
地点を選ぶ
項目を選ぶ
期間を選ぶ
表示オプションを選ぶ
他の都道府県
新潟県全地点
項目
気温
降水
日照/日射
積雪/降雪
風
湿度/気圧
雲量/天気
月平均気温
日最高気温の月平均
日最低気温の月平均
月最高気温
月最低気温
日最高気温の月最低※
日最低気温の月最高※
日平均気温 25 ℃以上の日数(月)
日平均気温 0 ℃未満の日数(月)
日最高気温 25 ℃以上の日数(月)
日最高気温 0 ℃未満の日数(月)
日最低気温 25 ℃以上の日数(月)
日最低気温 0 ℃未満の日数(月)
※官署(気象台等)のみ値があります
福島県
栃木県
富山県
地域の統計データを入手しよう
長岡技術科学大学
Nagaoka University of Technology
Iwahashi
53


今後、温暖化はどうなる？
新潟市
100年後は
15.4℃
+0.9
今は14.5℃
気温(℃)
年
直線に回帰
気温＝0.0133×年 －12.72
新潟市
100年後は
18.7℃
+4.2
今は14.5℃
気温(℃)
年
2次の多項式に回帰
気温＝0.0001×年² －0.4705×年＋459.1
長岡技術科学大学
Nagaoka University of Technology
Iwahashi
54

# 戦後の経済成長と温暖化




直線に回帰

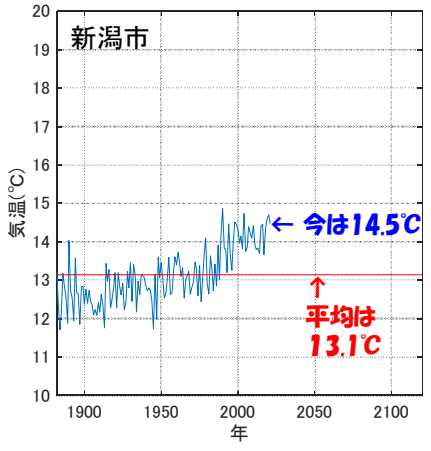



2次の多項式に回帰

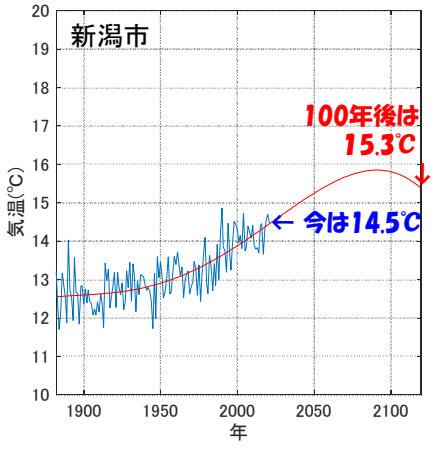
青木繁伸「統計数字を読み解くセンス」DOJIN選書



# 回帰多項式の次数 （慎重に選ぶこと）

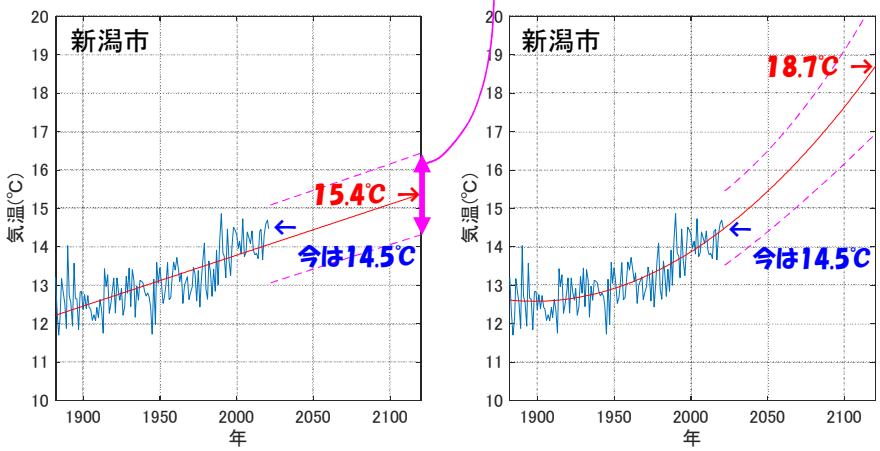



定数に回帰




4次の多項式に回帰

# 95% の確率でこの範囲内 区間推定




**直線に回帰**　　**2次の多項式に回帰**



# 標準偏差と予測区間（信頼度95%）

$$\hat{y}_k \pm t_{2.5,n-2}\sqrt{\left\{1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{n\,s_x^2}\right\}s_r^2}$$

**予測値　±ばらつく幅**

**回帰の誤差が大きいとばらつく幅も大きい**

$$\hat{y}_k = a + b\,x_k$$

$$a = \bar{y} - b\,\bar{x}$$

$$b = \frac{s_{xy}}{s_x^2}$$

**回帰の誤差の大きさ**

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$\begin{cases} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \end{cases}$$

$$\begin{cases} s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \end{cases}$$

Rで始めるデータサイエンス②

# Rを使って線形回帰

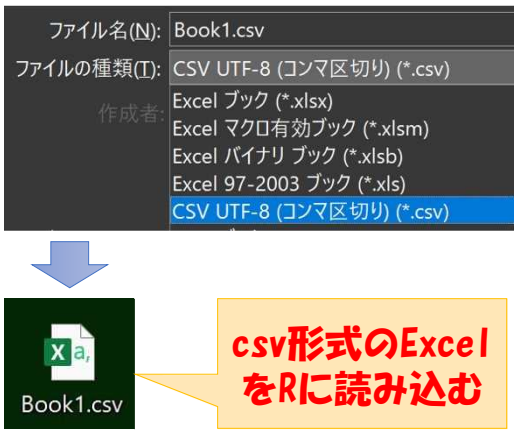
- 関係を数式で表す　～単回帰
- 単回帰の数理　～最小二乗法
- 単回帰の性質　～Rで単回帰
- 多項式に近似　～Yule-Walker
- 身近なデータに応用　～地球温暖化
- Excelデータ をRで処理



# Excelのデータは csv 形式で保存

①データを準備


②csv形式で保存




**csv形式のExcelをRに読み込む**

# Excel を R に読み込む

Excelがある
ディレクトリ

```
setwd("/Users/user/Desktop")
df <- read.csv("Book1.csv",header=T,row.names=1)
df
```

Excelの
ファイル名

1行目が
列の名前

1列目が
行の名前

```
   理科 数学
1    85   78
2    96   90
3    64   62
4    90   82
5    76   70
6    80   75
7    73   70
8    70   63
```



# 読み込んだデータを取り出す

2行目のデータ

```
df[2,]
 理科 数学
2   96   90
```

2列目のデータ

```
df[,2]
78 90 62 82 70 75 70 63
```

"理科"という名前
の列のデータ

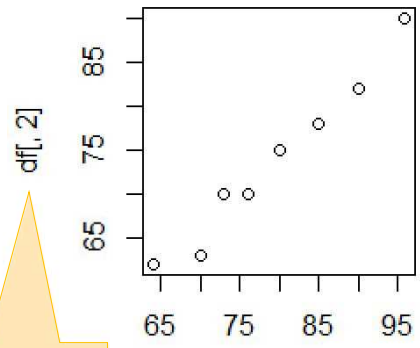
```
df$理科
85 96 64 90 76 80 73 70
```

"理科"が 70 未満
の行

```
subset(df, df$理科 <70)
3   64   62
```

# データの散布図をかく

> plot(df[,1],df[,2])　　☜Rに入力




☜Rの出力

**2列目のデータ**

**1列目のデータ**



初版：　　2022年8月

制作：　　岩橋政宏
所属：　　長岡技術科学大学