データの分析と知識発見

Introduction to Data Analysis

今回の構成

重回帰分析の係数の導入過程について学ぶ

重回帰分析における決定係数とその役割を学ぶ

Rを用いての回帰分析
その結果をグラフに表す方法を学ぶ

(重) 回帰分析

(重)回帰分析

$$y = a_1x_1 + a_2x_2 + \cdots + a_rx_r + b = \sum_{i=1}^{r} a_ix_i + b$$

y：目的変数

$x_i\,(i=1, 2, \cdots, r)$：説明変数

…

$(x_1^{(1)}, x_2^{(1)}, \cdots, x_r^{(1)}, y^{(1)})$

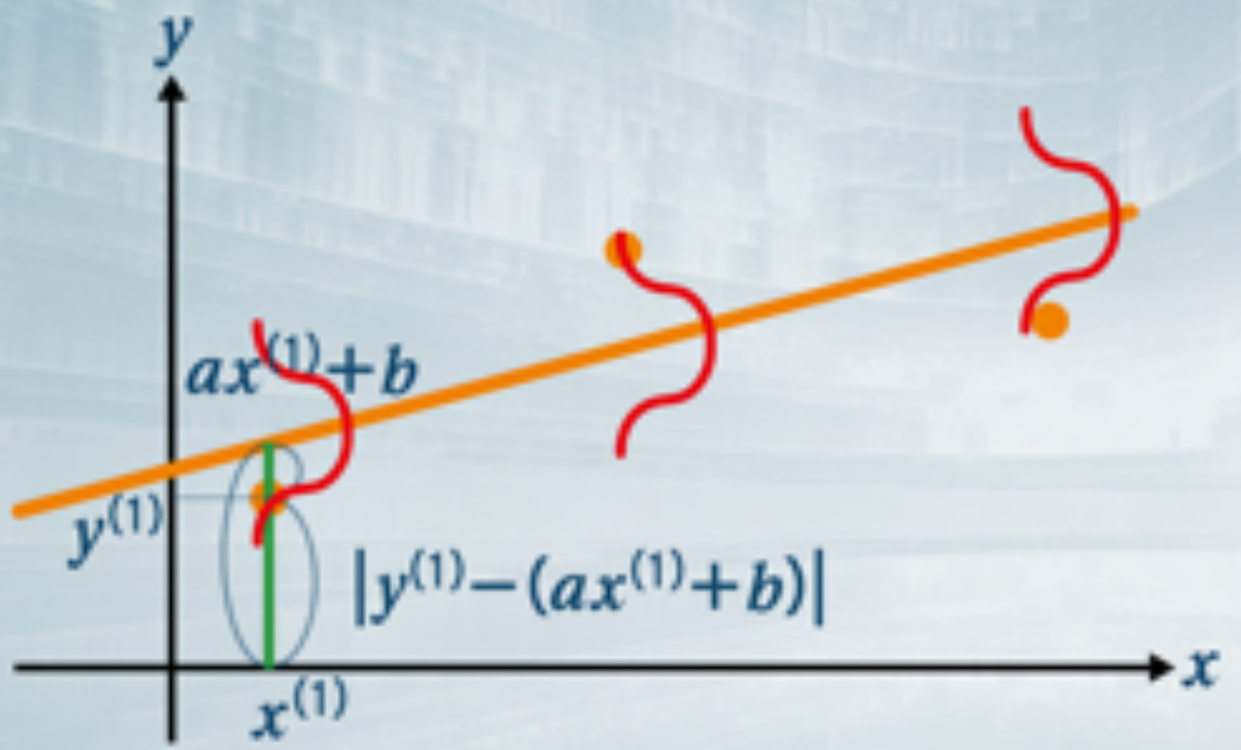
$(x_1^{(2)}, x_2^{(2)}, \cdots, x_r^{(2)}, y^{(2)})$

$(x_1^{(N)}, x_2^{(N)}, \cdots, x_r^{(N)}, y^{(N)})$

から偏回帰係数 a_i, b を求める

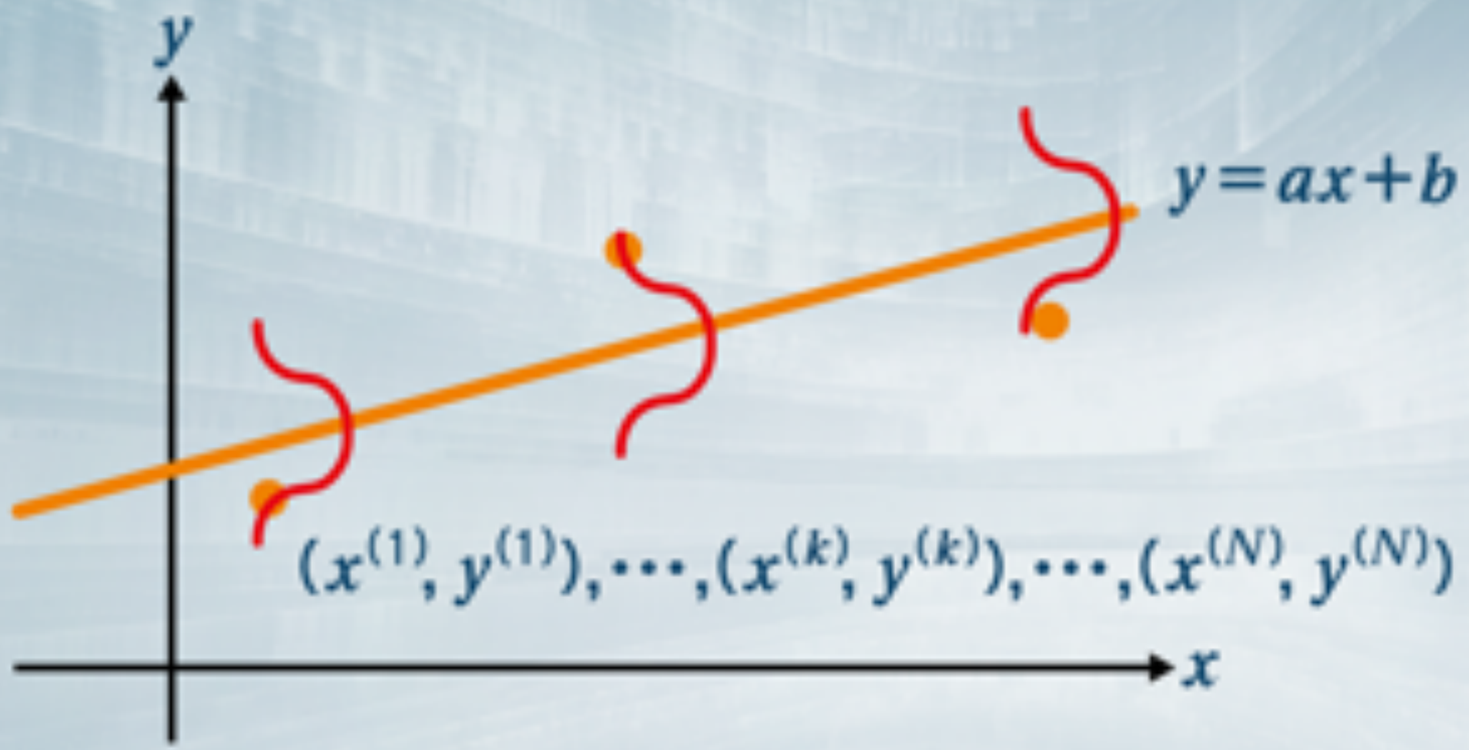
	x_1	x_2	$\cdots$	x_r	y
1	$x_1^{(1)}$	$x_2^{(1)}$	$\cdots$	$x_r^{(1)}$	$y^{(1)}$
2	$x_1^{(2)}$	$x_2^{(2)}$	$\cdots$	$x_r^{(2)}$	$y^{(2)}$
			$\vdots$		
N	$x_1^{(N)}$	$x_2^{(N)}$	$\cdots$	$x_r^{(N)}$	$y^{(N)}$

係数の導出

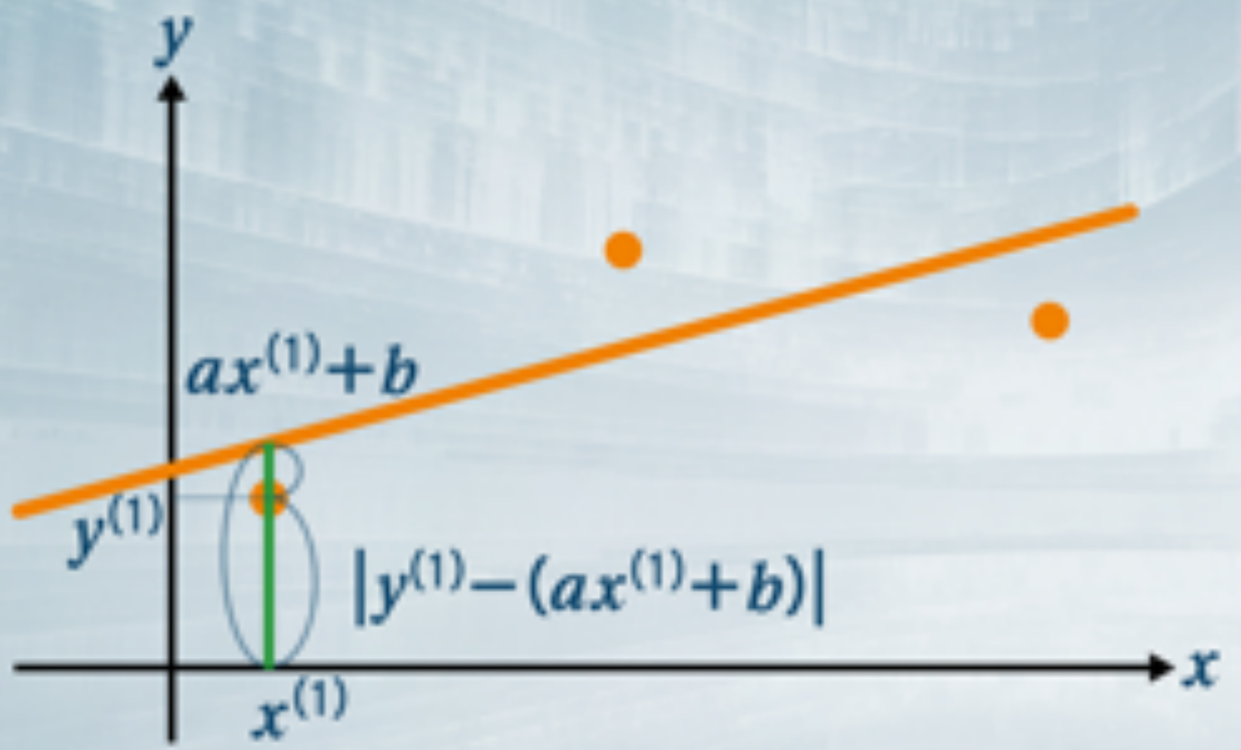


$$E=(y^{(1)}-(ax^{(1)}+b))^2$$

係数の導出

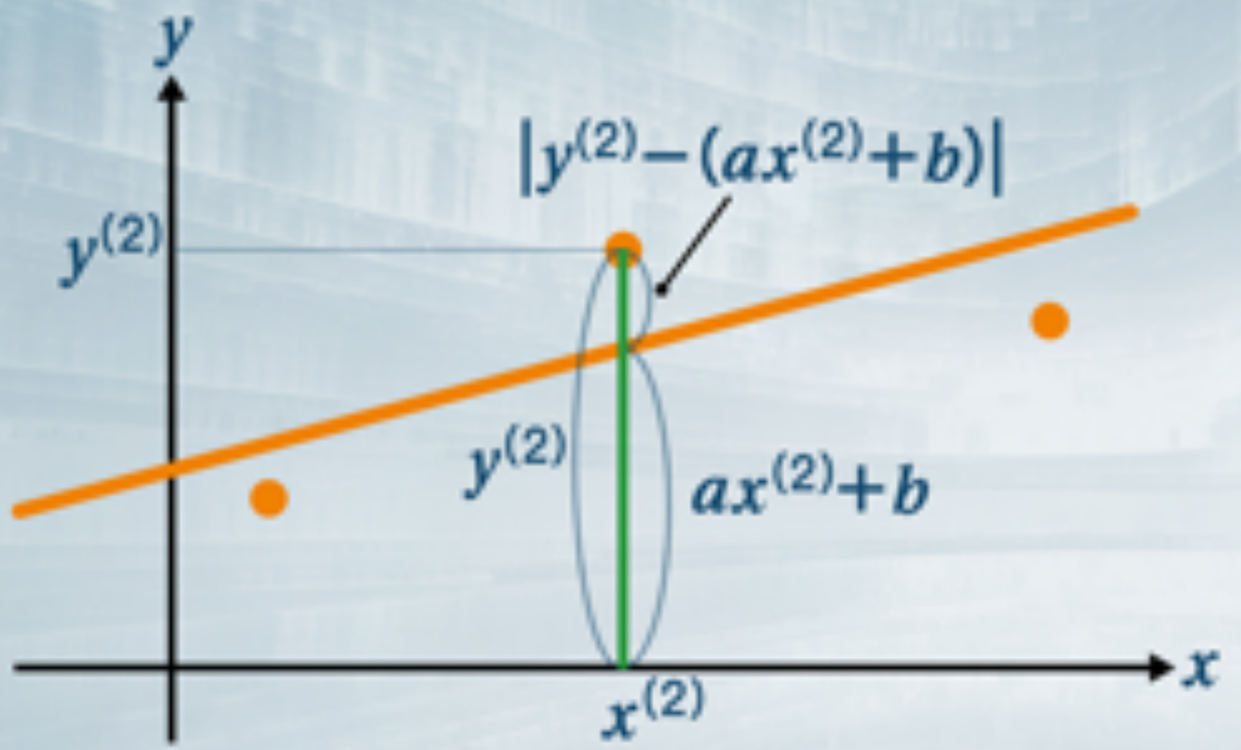


係数の導出



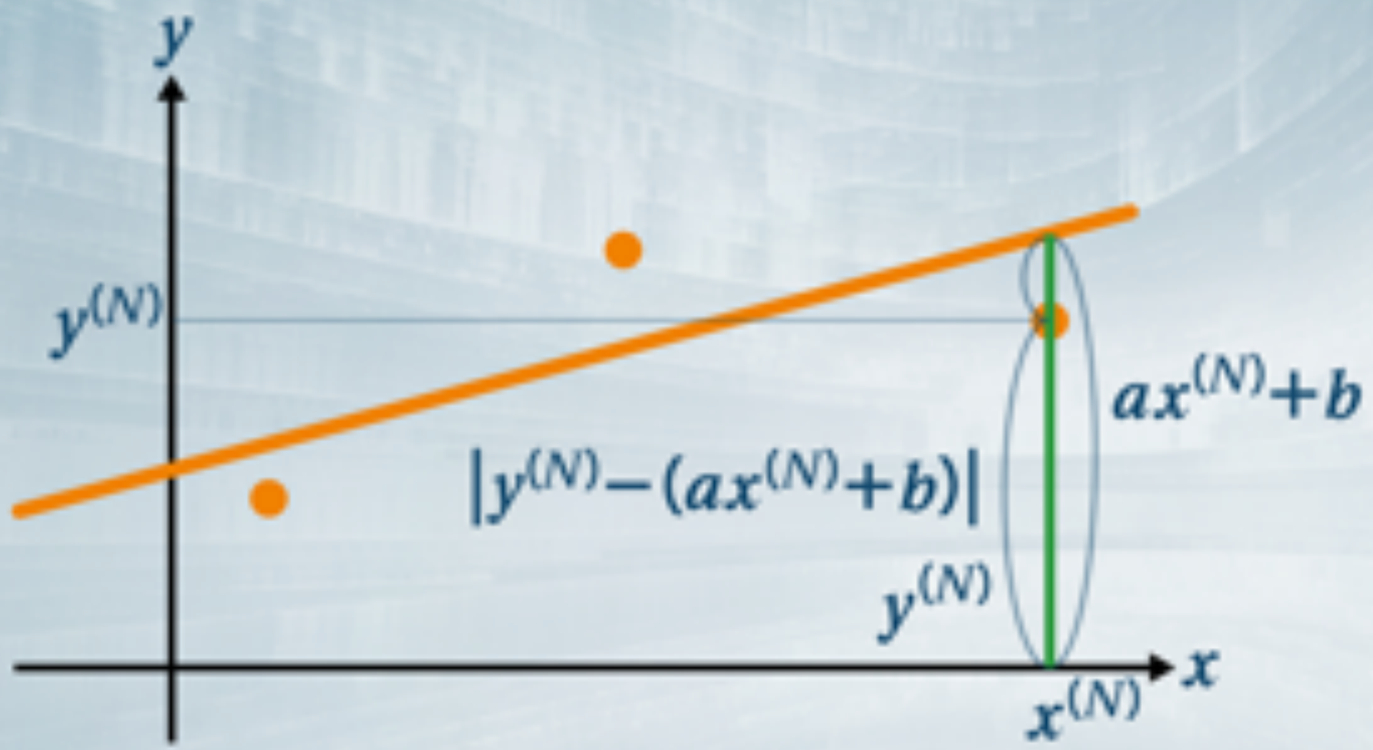
$$E=(y^{(1)}-(ax^{(1)}+b))^2$$

係数の導出



$$E=(y^{(1)}-(ax^{(1)}+b))^2$$
$$+(y^{(2)}-(ax^{(2)}+b))^2$$

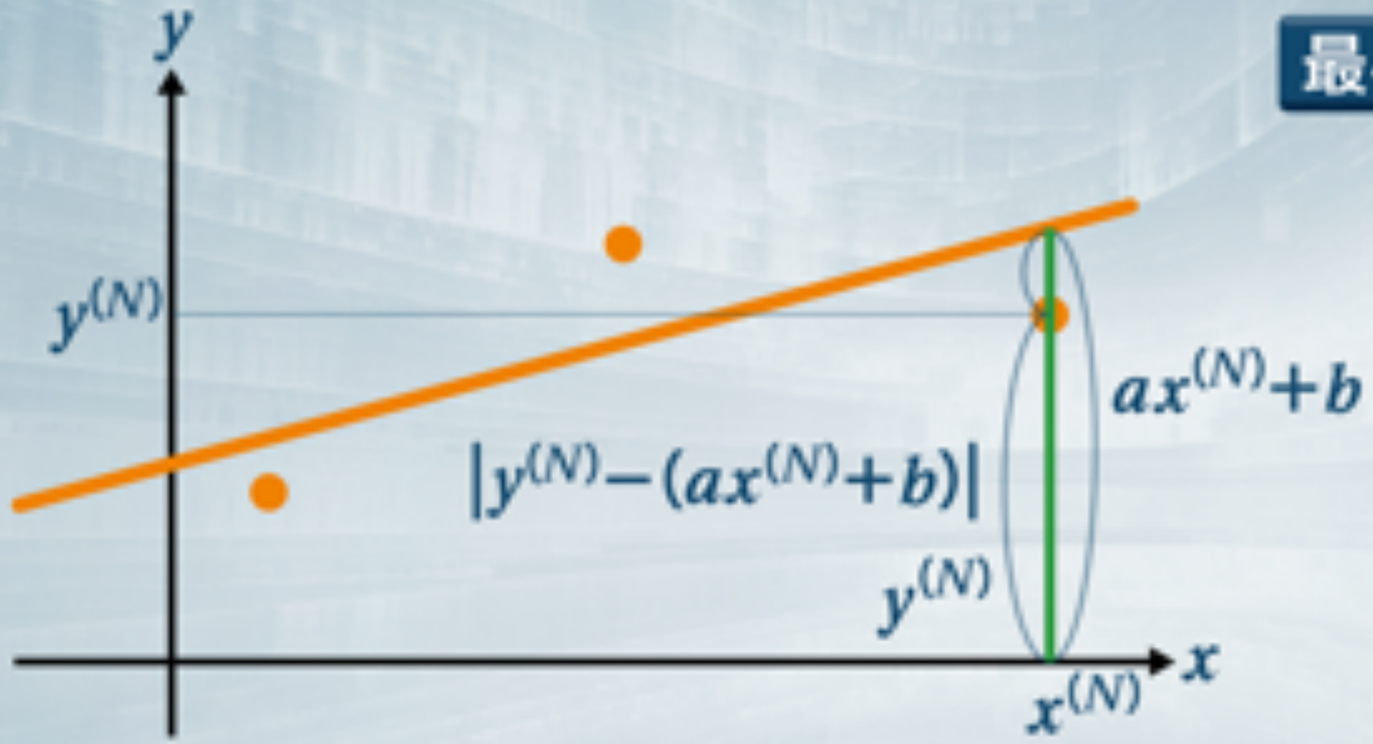
係数の導出



$$
\begin{aligned}
E = &(y^{(1)}-(ax^{(1)}+b))^2 \\
&+(y^{(2)}-(ax^{(2)}+b))^2 \\
&\quad\vdots \\
&+(y^{(N)}-(ax^{(N)}+b))^2
\end{aligned}
$$

Eが最小となるa, bを求める

係数の導出



最小2乗法

$$
\begin{aligned}
E = &(y^{(1)}-(ax^{(1)}+b))^2 \\
&+(y^{(2)}-(ax^{(2)}+b))^2 \\
&\vdots \\
&+(y^{(N)}-(ax^{(N)}+b))^2
\end{aligned}
$$

Eが最小となるa, bを求める

回帰係数の導出

最小2乗法

$$E=(y^{(1)}-(ax^{(1)}+b))^2$$
$$+(y^{(2)}-(ax^{(2)}+b))^2$$
$$+\cdots+(y^{(N)}-(ax^{(N)}+b))^2$$
$$=\sum_{p=1}^{N}(y^{(p)}-(ax^{(p)}+b))^2$$

Eが最小となるa, bを求める

a, bで偏微分すると

$$\sum_{p=1}^{N}(y^{(p)}-(ax^{(p)}+b))x^{(p)}=0$$

$$\sum_{p=1}^{N}(y^{(p)}-(ax^{(p)}+b))=0$$

回帰係数の導出

$$\sum_{p=1}^{N}(y^{(p)}-(ax^{(p)}+b))=0$$

$$\frac{1}{N}\sum_{p=1}^{N}y^{(p)}=a\cdot\frac{1}{N}\sum_{p=1}^{N}x^{(p)}+\frac{1}{N}\sum_{p=1}^{N}b$$

$$\rightarrow \quad \mu_y=a\mu_x+b$$

回帰係数の導出

$$\sum_{p=1}^{N}(y^{(p)}-(ax^{(p)}+b))=0$$

$$\frac{1}{N}\sum_{p=1}^{N}y^{(p)}=a\cdot\frac{1}{N}\sum_{p=1}^{N}x^{(p)}+\frac{1}{N}\sum_{p=1}^{N}b$$

$$\rightarrow \quad \mu_y=a\mu_x+b$$

$$y=ax+b$$

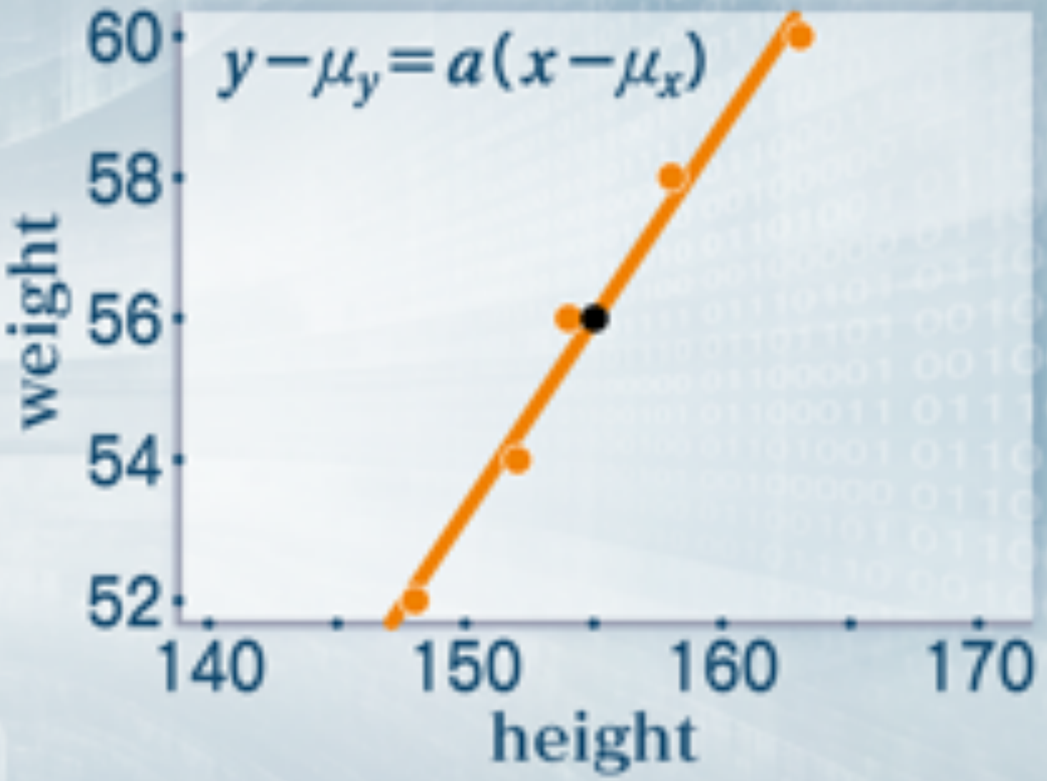
回帰係数の導出

$$\sum_{p=1}^{N}(y^{(p)}-(ax^{(p)}+b))=0$$

$$\frac{1}{N}\sum_{p=1}^{N}y^{(p)}=a\cdot\frac{1}{N}\sum_{p=1}^{N}x^{(p)}+\frac{1}{N}\sum_{p=1}^{N}b$$

$$\rightarrow \quad \mu_y=a\mu_x+b$$

回帰直線 $y=ax+b$ は
それぞれの中心(平均)の点を通る



回帰係数の導出

$y-\mu_y=a(x-\mu_x)$

回帰直線はそれぞれの中心を通る

$(x^{(1)}, y^{(1)}), \cdots, (x^{(p)}, y^{(p)}),$
$\cdots, (x^{(N)}, y^{(N)})$

中心化

$x^{(p)} \Longleftarrow (x^{(p)}-\mu_x)$
$y^{(p)} \Longleftarrow (y^{(p)}-\mu_y)$

$$E=(y^{(1)}-ax^{(1)})^2+(y^{(2)}-ax^{(2)})^2$$
$$+\cdots+(y^{(N)}-ax^{(N)})^2$$
$$=\sum_{p=1}^{N}(y^{(p)}-ax^{(p)})^2$$

$$\sum_{p=1}^{N}(y^{(p)}-ax^{(p)})x^{(p)}=0$$

N（または$N-1$）で割ると
分散共分散を用いて

$$s_{xy}-as_{xx}=0$$

回帰係数の導出

条件を整理すると

$$s_{xy} - a s_{xx} = 0$$

$$\mu_y = a\mu_x + b$$

$$\boxed{\begin{aligned} s_{xy} &= a s_{xx} \\ b &= u_y - a\mu_x \end{aligned}}$$

$s_{xx} \neq 0$ のとき

$$\boxed{a = \frac{s_{xy}}{s_{xx}},\ b = -\frac{s_{xy}}{s_{xx}}\mu_x + \mu_y}$$

r 次元の場合

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_r x_r + b = \sum_{i=1}^{r} a_i x_i + b$$

同様に計算すると・・・

$$\boxed{\begin{aligned} s_{1y} &= s_{11} a_1 + s_{12} a_2 + \cdots + s_{1r} a_r \\ s_{2y} &= s_{21} a_1 + s_{22} a_2 + \cdots + s_{2r} a_r \\ &\vdots \\ s_{ry} &= s_{r1} a_1 + s_{r2} a_2 + \cdots + s_{rr} a_r \\ b &= \mu_y - (a_1\mu_1 + a_2\mu_2 + \cdots + a_r\mu_r) \end{aligned}}$$

回帰係数の計算

最小2乗法にて導出

bに関して

2次元の場合

$$y = ax + b$$
$$\mu_y = a\mu_x + b$$
$$y - \mu_y = a(x - \mu_x)$$

r次元の場合

$$y = a_1 x_1 + \cdots + a_r x_r + b$$
$$\mu_y = a_1 \mu_1 + \cdots + a_r \mu_r + b$$
$$y - \mu_r = a_1(x_1 - \mu_1) + \cdots + a_r(x_r - \mu_r)$$

回帰係数の計算

最小2乗法にて導出

bに関して

2次元の場合

$$y = ax + b$$
$$\mu_y = a\mu_x + b$$
$$y - \mu_y = a(x - \mu_x)$$

r次元の場合

$$y = a_1 x_1 + \cdots + a_r x_r + b$$
$$\mu_y = a_1 \mu_1 + \cdots + a_r \mu_r + b$$
$$y - \mu_r = a_1(x_1 - \mu_1) + \cdots + a_r(x_r - \mu_r)$$

a_iに関して

2次元の場合

$$s_{xy} = a s_{xx}$$

r次元の場合

$$\begin{pmatrix} s_{1y} \\ \vdots \\ s_{ry} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1r} \\ s_{21} & s_{22} & \cdots & s_{2r} \\ \vdots & & \ddots & \vdots \\ s_{r1} & s_{r2} & \cdots & s_{rr} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_r \end{pmatrix}$$

回帰係数の計算

最小2乗法にて導出

bに関して

2次元の場合

$$y = ax + b$$
$$\mu_y = a\mu_x + b$$
$$y - \mu_y = a(x - \mu_x)$$

r次元の場合

$$y = a_1 x_1 + \cdots + a_r x_r + b$$
$$\mu_y = a_1 \mu_1 + \cdots + a_r \mu_r + b$$
$$y - \mu_r = a_1(x_1 - \mu_1) + \cdots + a_r(x_r - \mu_r)$$

回帰直線は中心を通る

a_iに関して

2次元の場合

$$s_{xy} = a s_{xx}$$

r次元の場合

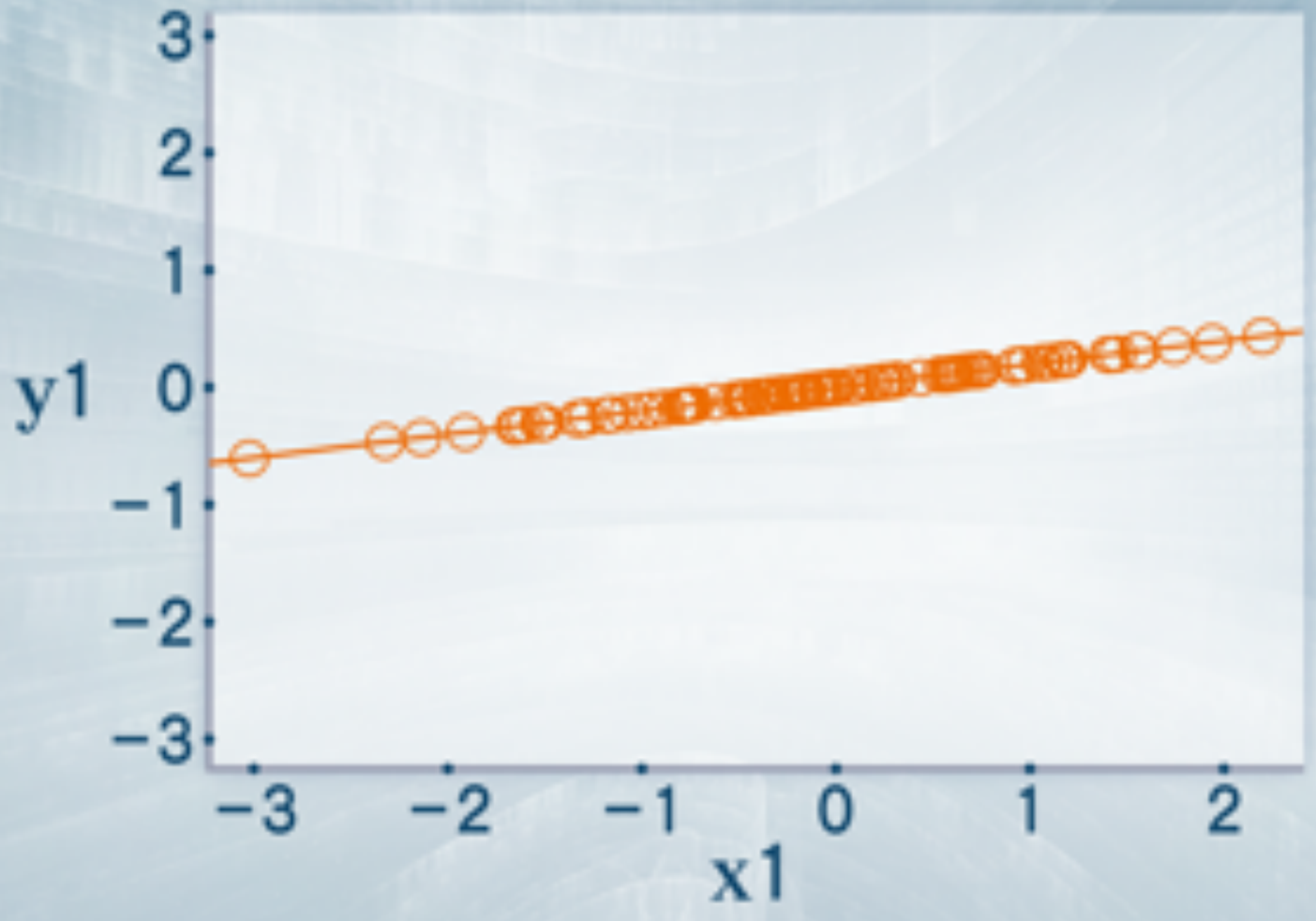
$$\begin{pmatrix} s_{1y} \\ \vdots \\ s_{ry} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1r} \\ s_{21} & s_{22} & \cdots & s_{2r} \\ \vdots & & \ddots & \vdots \\ s_{r1} & s_{r2} & \cdots & s_{rr} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_r \end{pmatrix}$$

逆行列があれば係数が導出できる

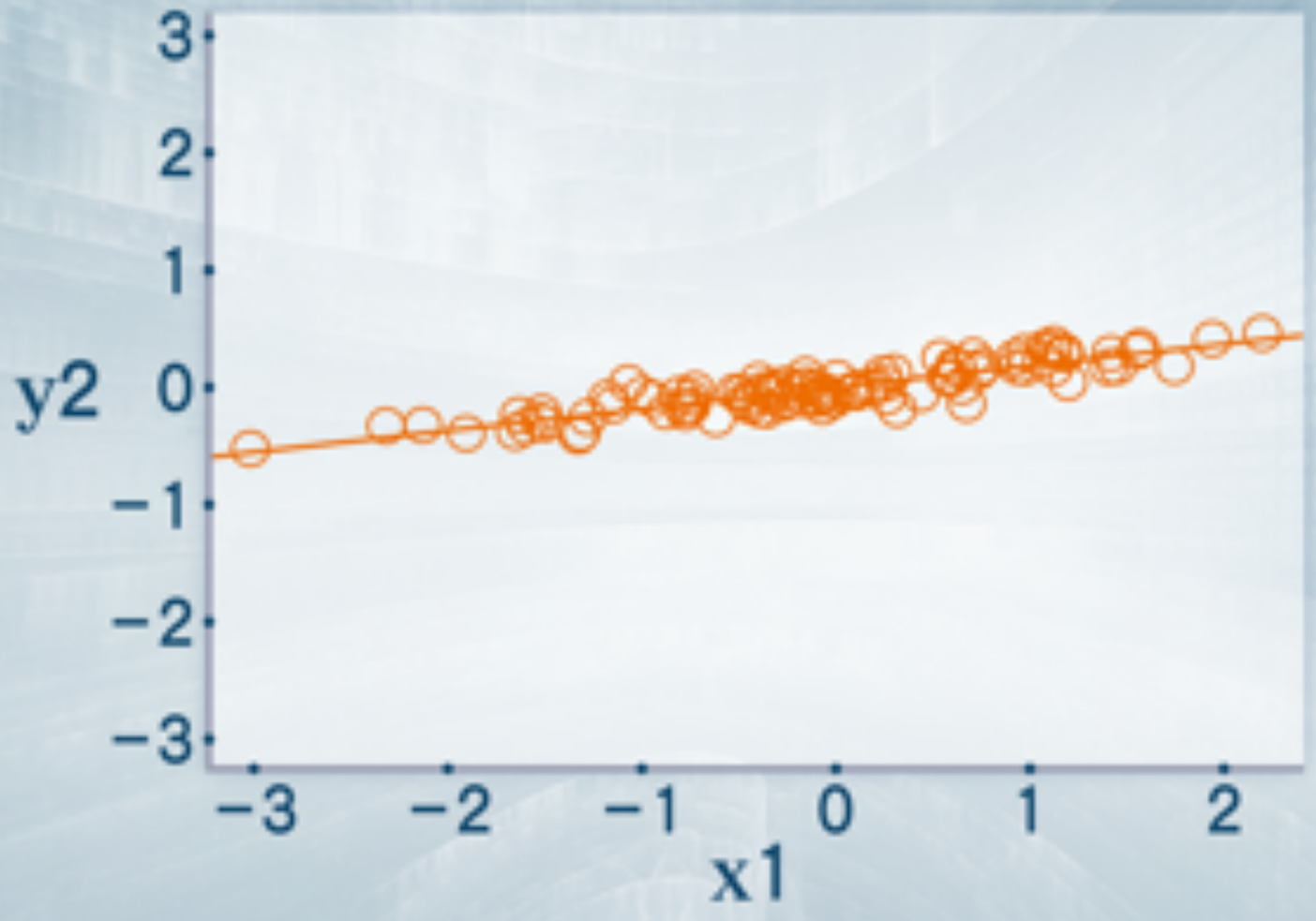
データの分析と知識発見

Introduction to Data Analysis

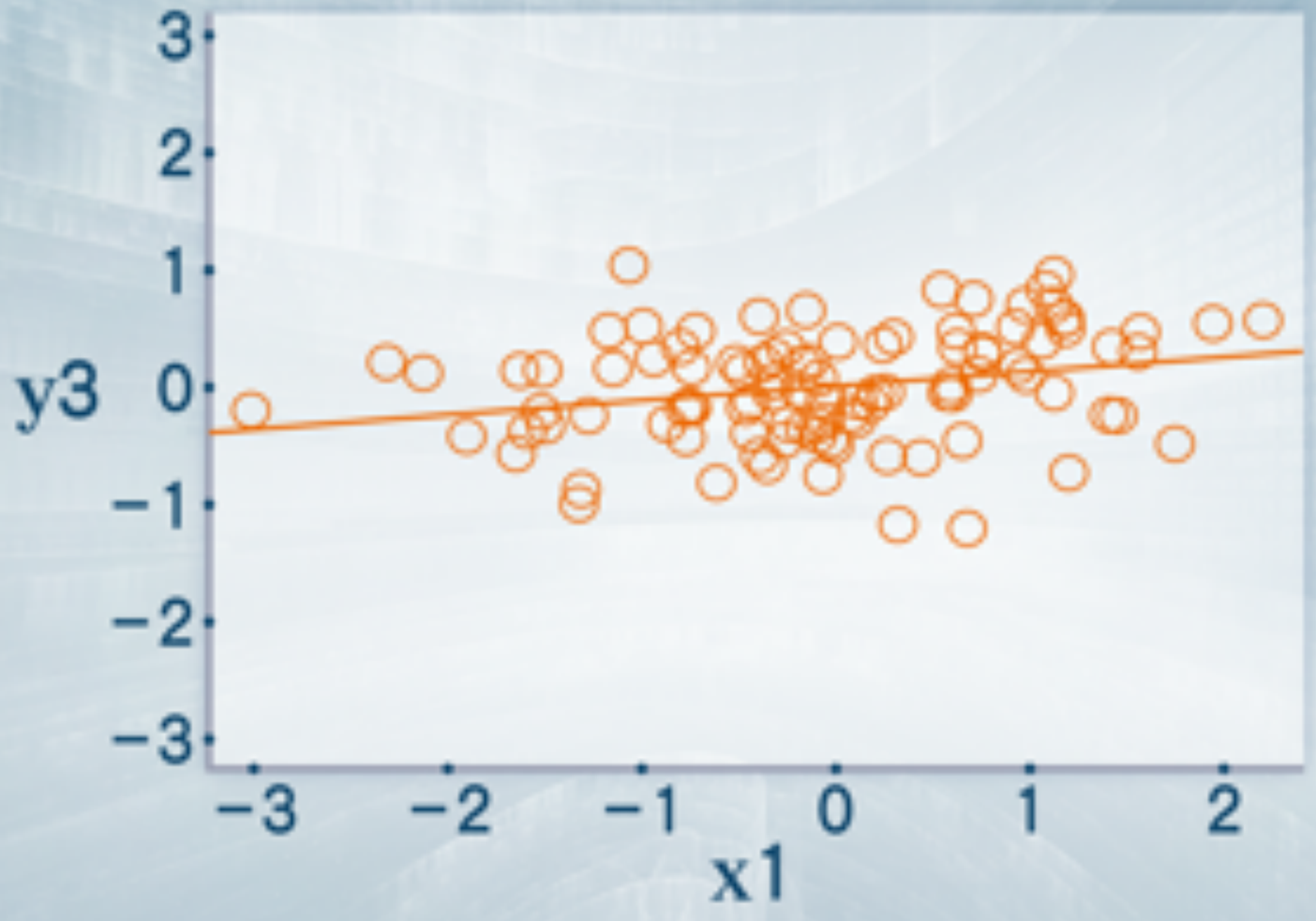
予測の精度



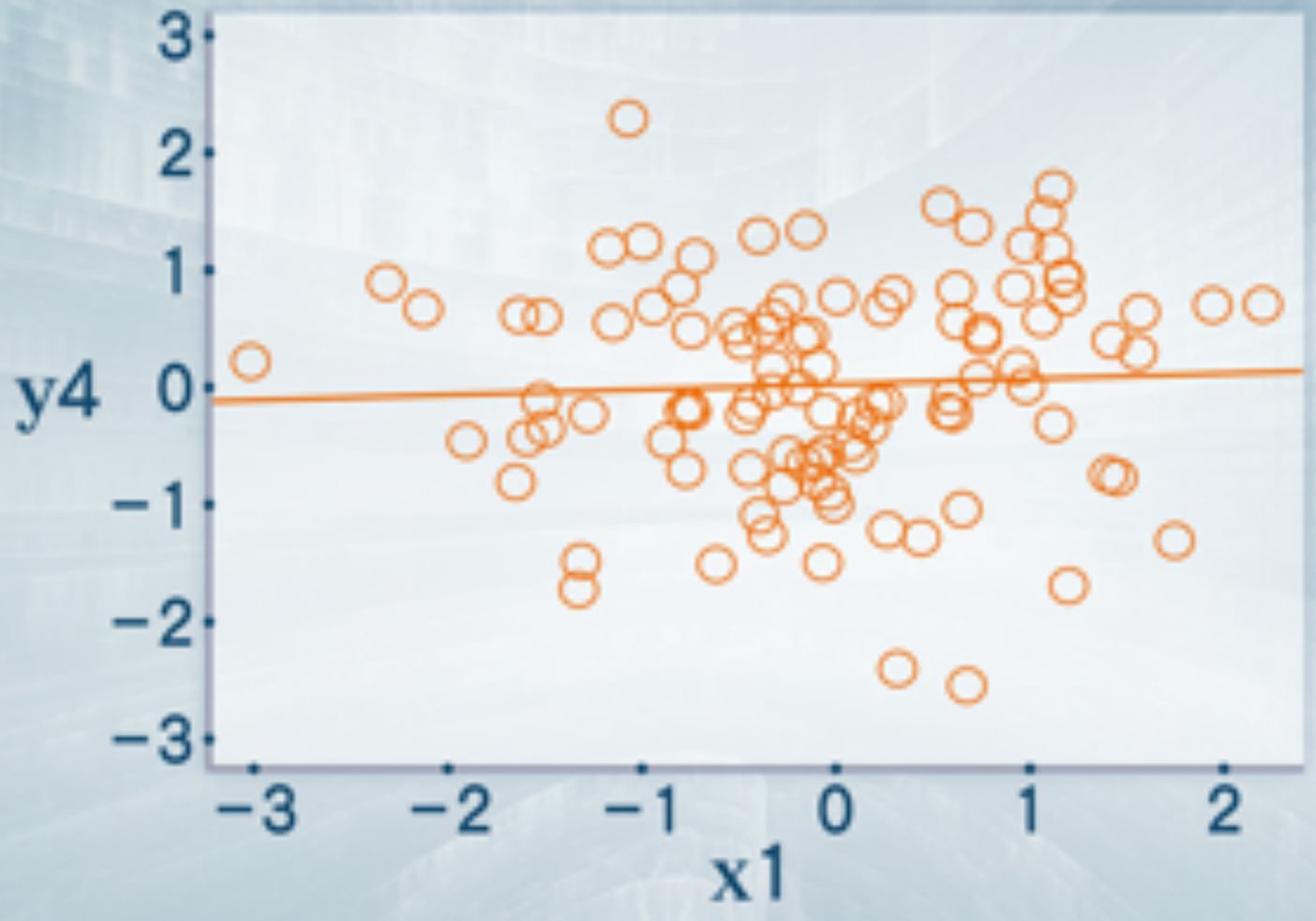
予測の精度



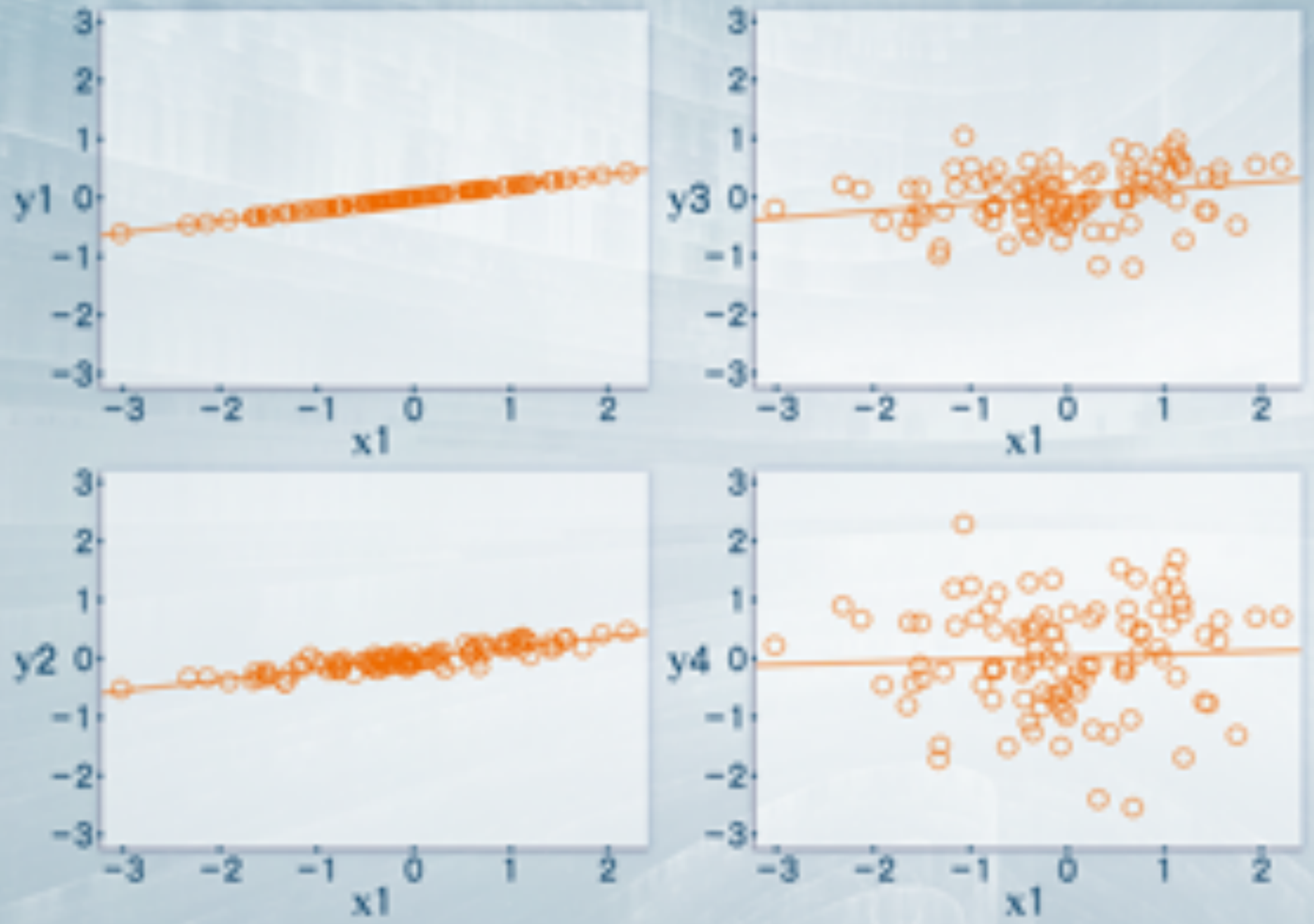
予測の精度



予測の精度



予測の精度



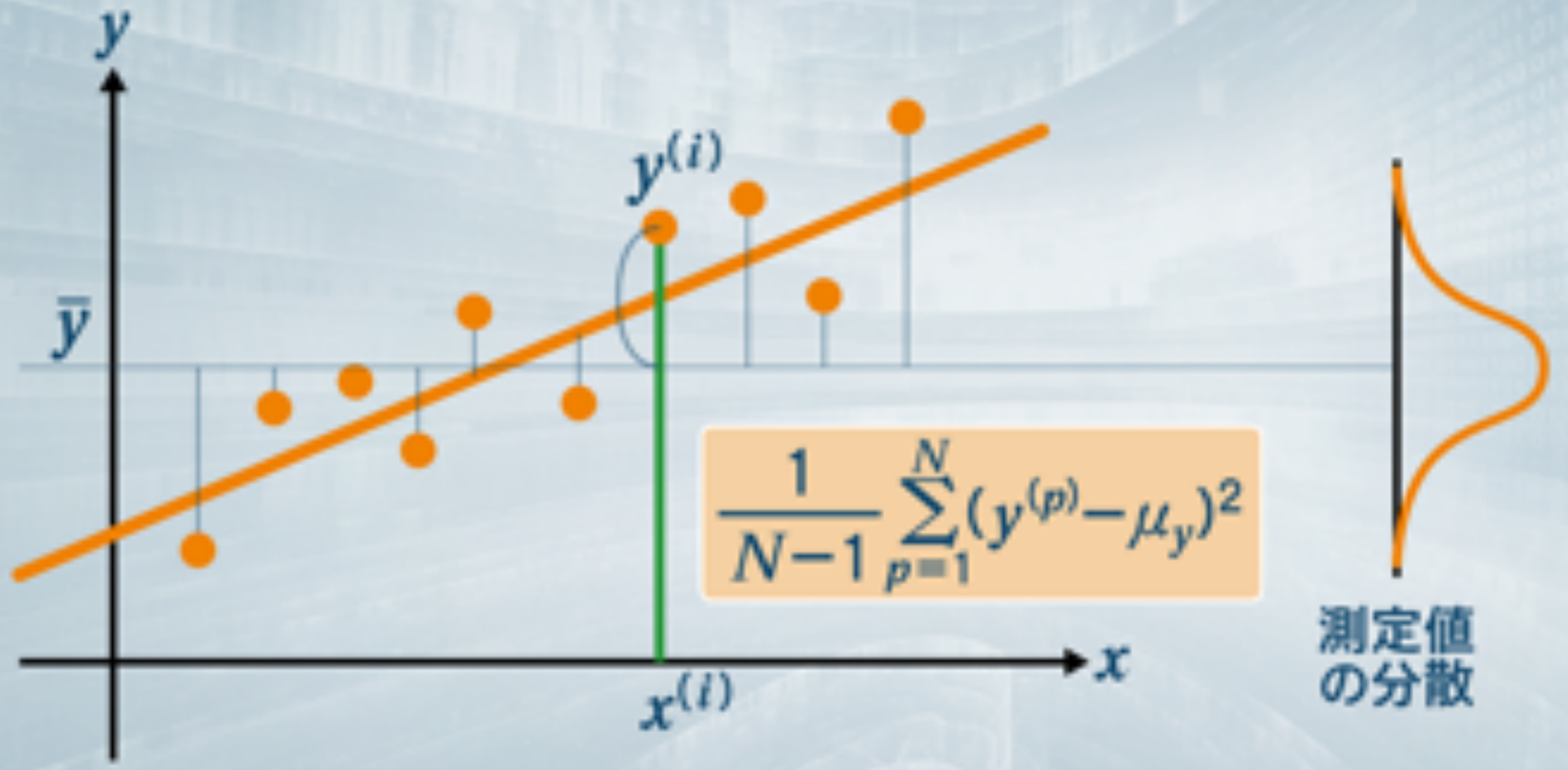
$y = 0.2x + \xi$

（ξ：正規乱数）

$s_{xx} \neq 0$ であれば
係数を求めることは
できる

値が求まるからといって
特徴をよく表している
というわけではない

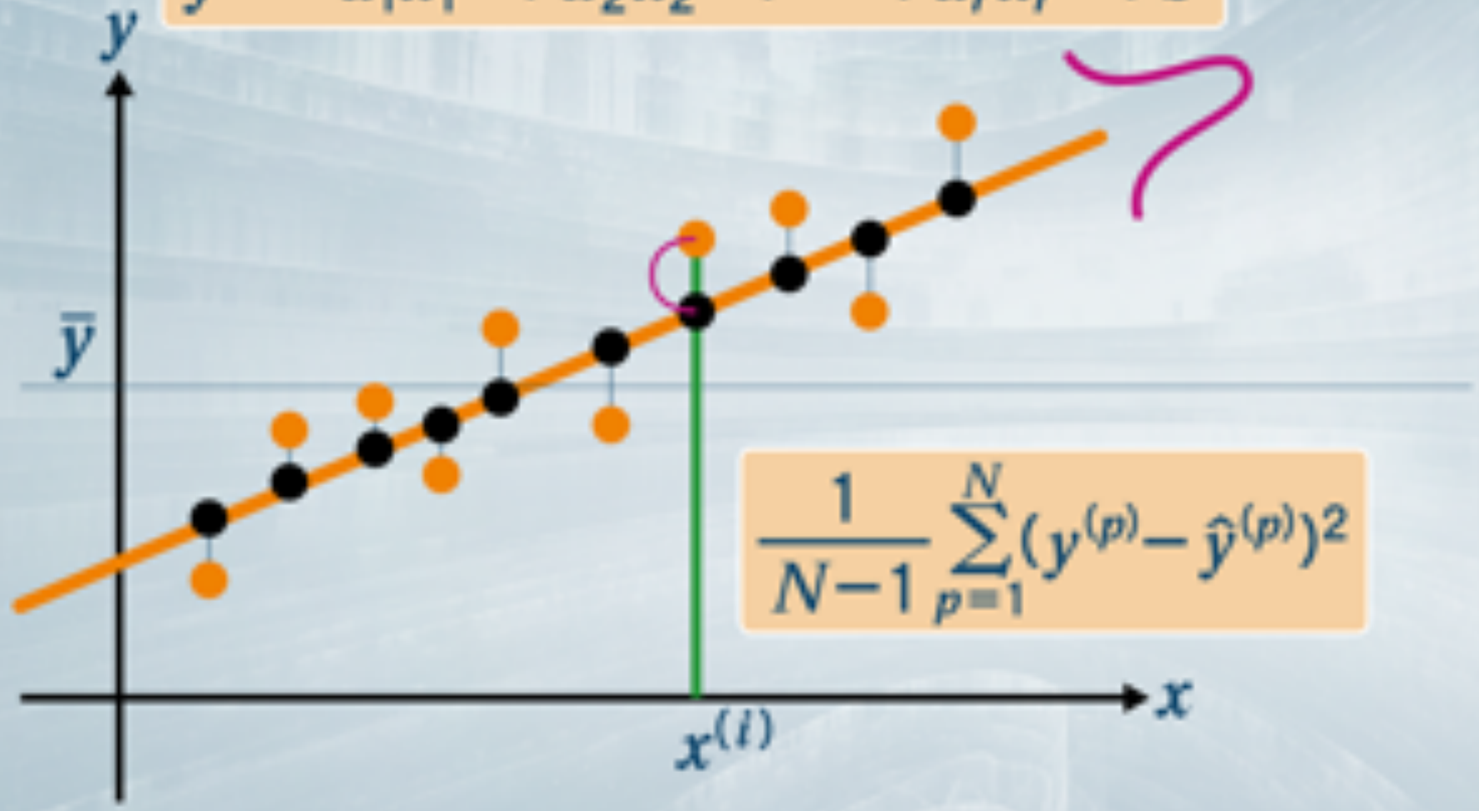
決定係数



$$\frac{1}{N-1}\sum_{p=1}^{N}(y^{(p)}-\mu_y)^2$$

決定係数

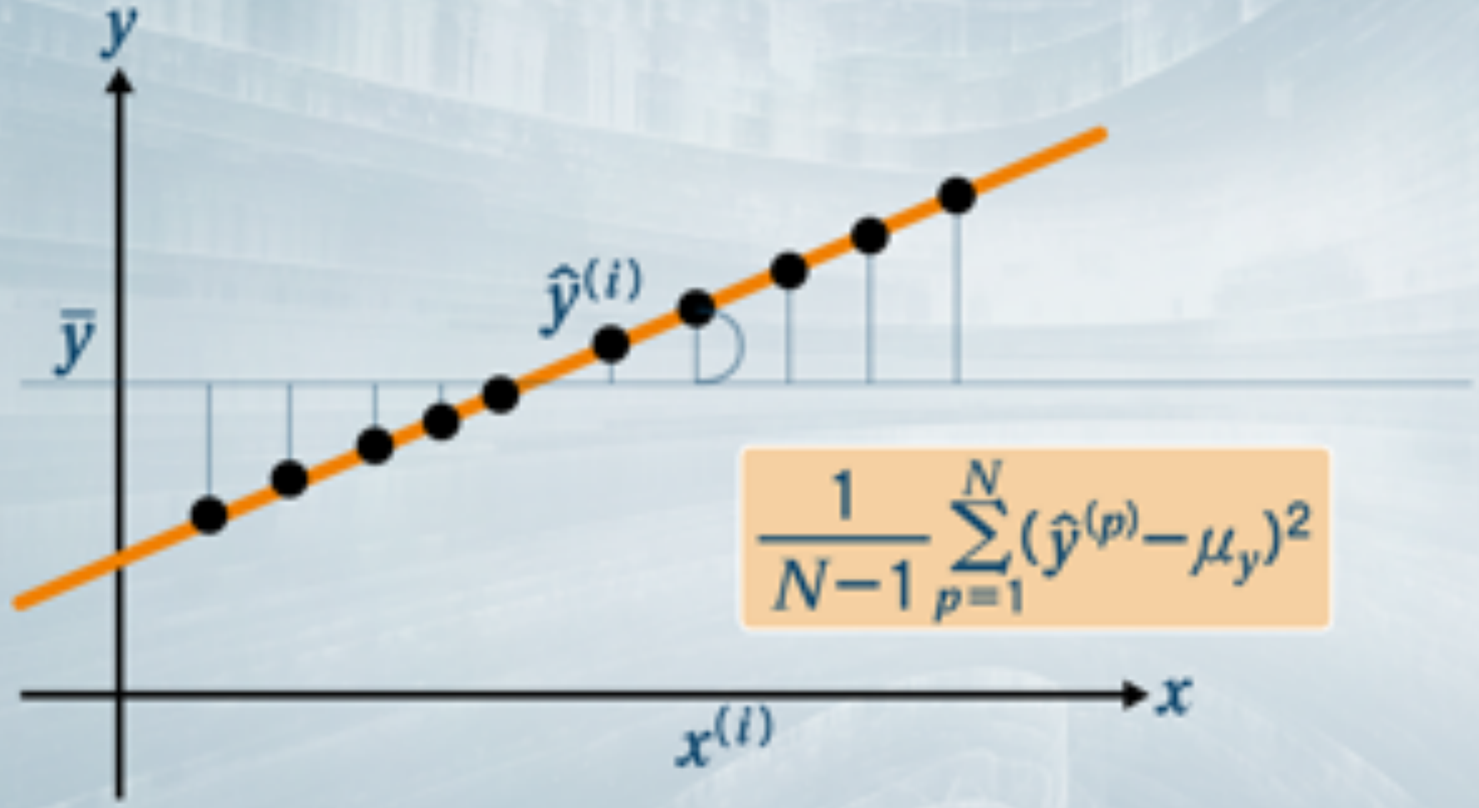
$$y^{(p)} = \hat{a}_1 x_1^{(p)} + \hat{a}_2 x_2^{(p)} + \cdots + \hat{a}_r x_r^{(p)} + \hat{b}$$



$$\frac{1}{N-1}\sum_{p=1}^{N}(y^{(p)} - \hat{y}^{(p)})^2$$

残差
の分散

決定係数



$$\frac{1}{N-1}\sum_{p=1}^{N}(\hat{y}^{(p)}-\mu_y)^2$$

予測値
の分散

偏相関係数

見かけの相関

x_i と x_j の直接の関係だけでなく
$x_1, x_2, \cdots x_r$ を通して相関係数に影響を与えることもある

$$x_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ir}x_r + b_i + e_i$$
$$x_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jr}x_r + b_j + e_j$$

e_i と e_j との相関 ⟶ 偏相関係数

Rによる計算

```
> h1 <- read.table ("ch071.dat", header=T, row.names=1)
> h1
```

```
> h2 <- lm(weight ~ height, data=h1)
> h2
```

```
> plot(h1, pch=16)
> abline(h2)
```

Rによる計算

```
> h1 <- read.table ("ch071.dat", header=T, row.names=1)
> h1
```

name	height	weight
A1	147.3	52.3
A2	149.9	53.2
A3	152.4	54.5
A4	154.9	55.9
A5	157.5	57.3
A6	160.0	58.6
A7	162.6	60.0
A8	165.1	61.4
A9	167.6	63.2
A10	170.2	64.5
A11	172.7	66.4
A12	175.3	68.2
A13	177.8	70.0
A14	180.3	72.3
A15	182.9	74.5

```
> h2 <- lm(weight ~ height, data=h1)
> h2
```

```
> plot(h1, pch=16)
> abline(h2)
```

Rによる計算

```
> h1 <- read.table ("ch071.dat", header=T, row.names=1)
> h1
```

```
> h2 <- lm(weight ~ height, data=h1)
> h2
```

lm（目的変数 ～ 説明変数, data ＝データ名）
複数の場合は
lm（目的変数 ～ 説明変数1＋説明変数2＋…, data ＝データ名）

Rによる計算

```
> h1 <- read.table ("ch071.dat", header=T, row.names=1)
> h1
```

```
> h2 <- lm(weight ~ height, data=h1)
> h2
```

```
> plot(h1, pch=16)
> abline(h2)
```

決定係数

$$y^{(p)} = \hat{y}^{(p)} + e^{(p)} = \hat{a}x^{(p)} + \hat{b} + e^{(p)}$$

$$\hat{y}^{(p)} = \hat{a}x^{(p)} + \hat{b}$$

として、誤差の2乗を計算すると

$$S_{yy} = \frac{1}{N-1}\sum_{p=1}^{N}(y^{(p)} - \mu_y)^2$$

$$= \frac{1}{N-1}\sum_{p=1}^{N}(y^{(p)} - \hat{y}^{(p)})^2 + \frac{1}{N-1}\sum_{p=1}^{N}(\hat{y}^{(p)} - \mu_y)^2$$

$$\begin{pmatrix}\text{測定値}\\ \text{の分散}\end{pmatrix} = \begin{pmatrix}\text{残差}\\ \text{の分散}\end{pmatrix} + \begin{pmatrix}\text{予測値}\\ \text{の分散}\end{pmatrix}$$

決定係数

$$y^{(p)} = \hat{y}^{(p)} + e^{(p)} = \hat{a}x^{(p)} + \hat{b} + e^{(p)}$$

$$\hat{y}^{(p)} = \hat{a}x^{(p)} + \hat{b}$$

として、誤差の2乗を計算すると

$$S_{yy} = \frac{1}{N-1}\sum_{p=1}^{N}(y^{(p)} - \mu_y)^2$$

$$= \frac{1}{N-1}\sum_{p=1}^{N}(y^{(p)} - \hat{y}^{(p)})^2 + \frac{1}{N-1}\sum_{p=1}^{N}(\hat{y}^{(p)} - \mu_y)^2$$

$$\begin{pmatrix}\text{測定値}\\ \text{の分散}\end{pmatrix} = \begin{pmatrix}\text{残差}\\ \text{の分散}\end{pmatrix} + \begin{pmatrix}\text{予測値}\\ \text{の分散}\end{pmatrix}$$

決定係数

$$S_T = S_E + S_R$$

T：トータル (Total)
E：残差 (Error)
R：回帰 (Regression)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$$