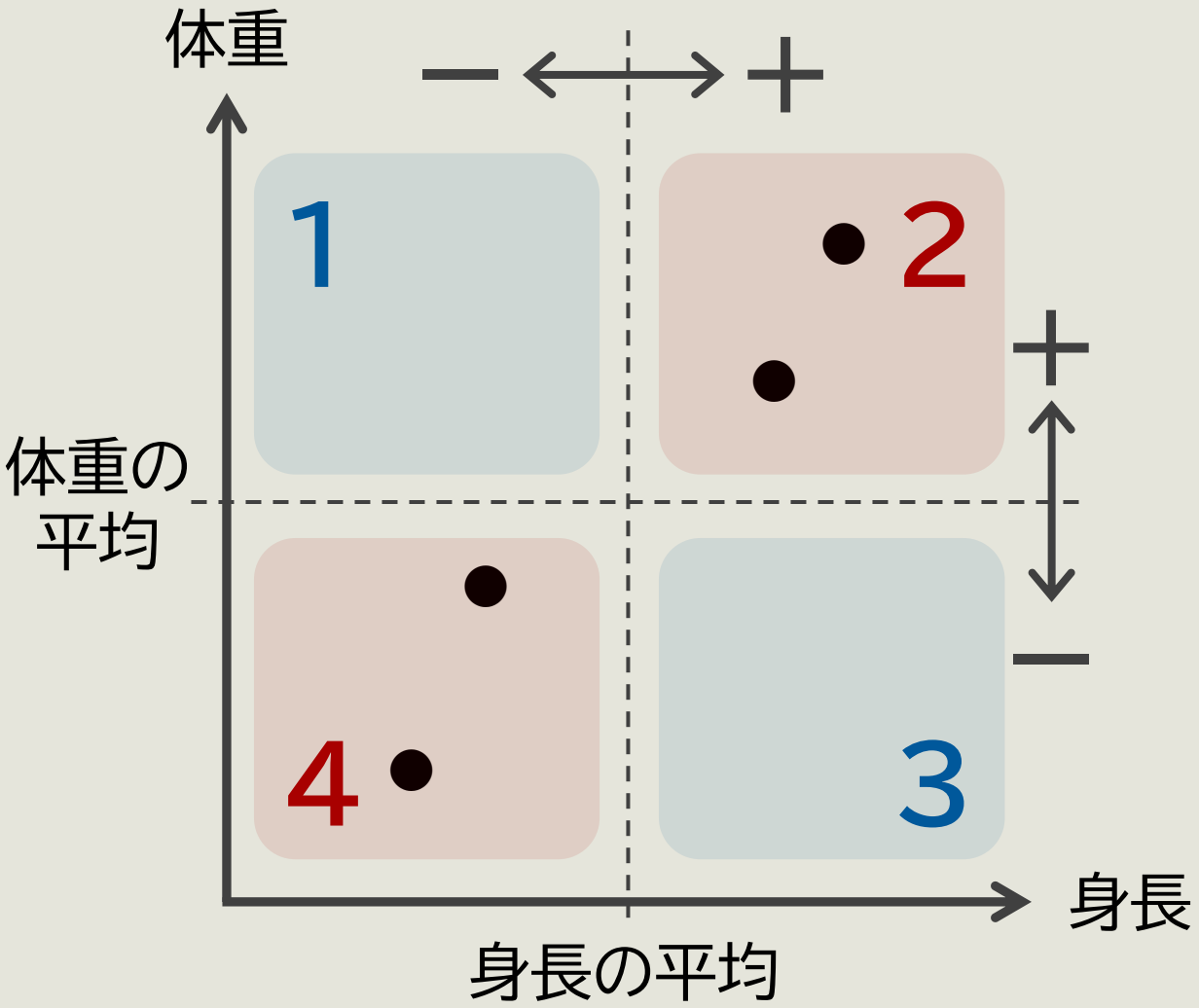
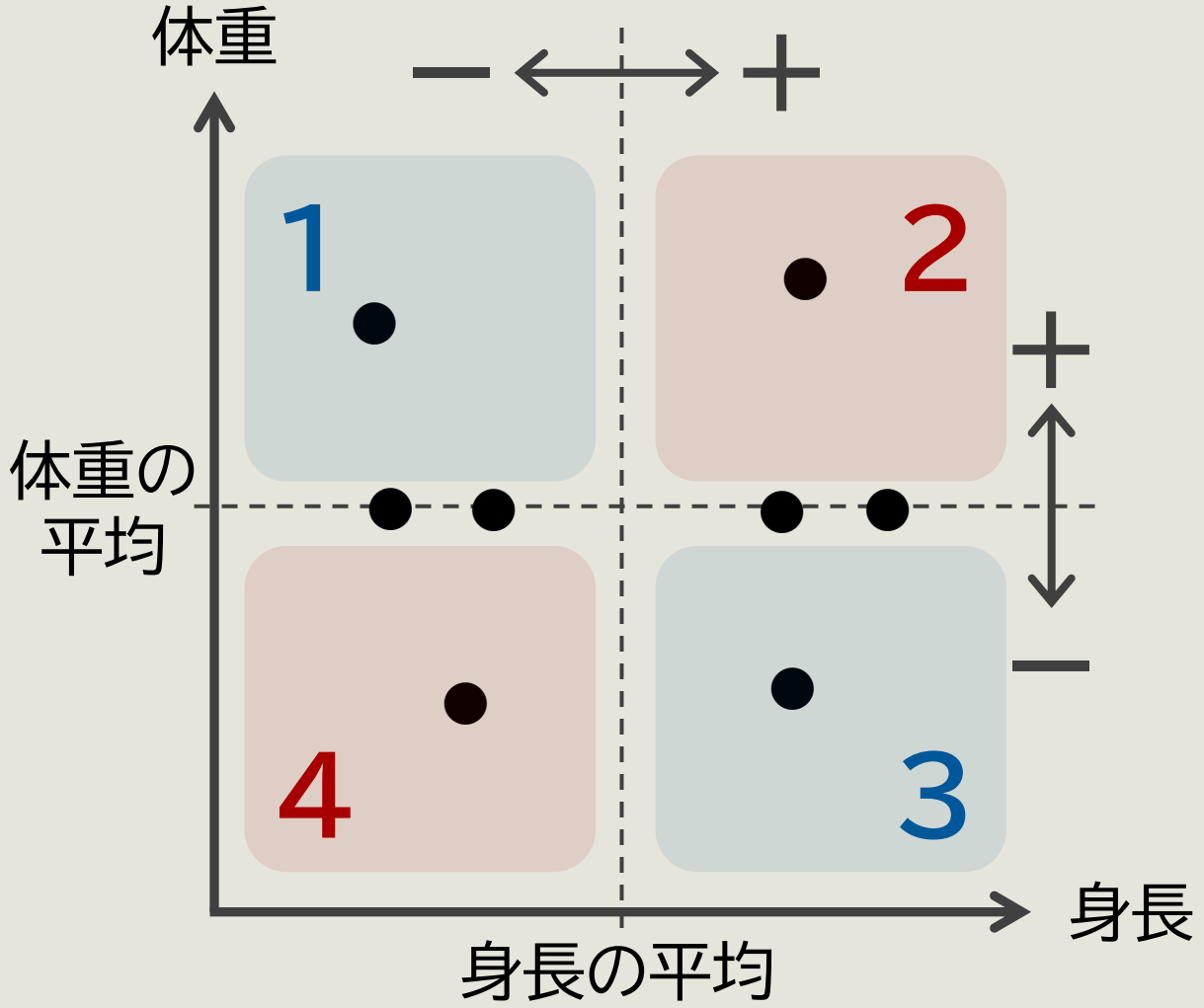
いちばん理解できる
統計学ベーシック講座
【相関分析・回帰分析】

配布資料

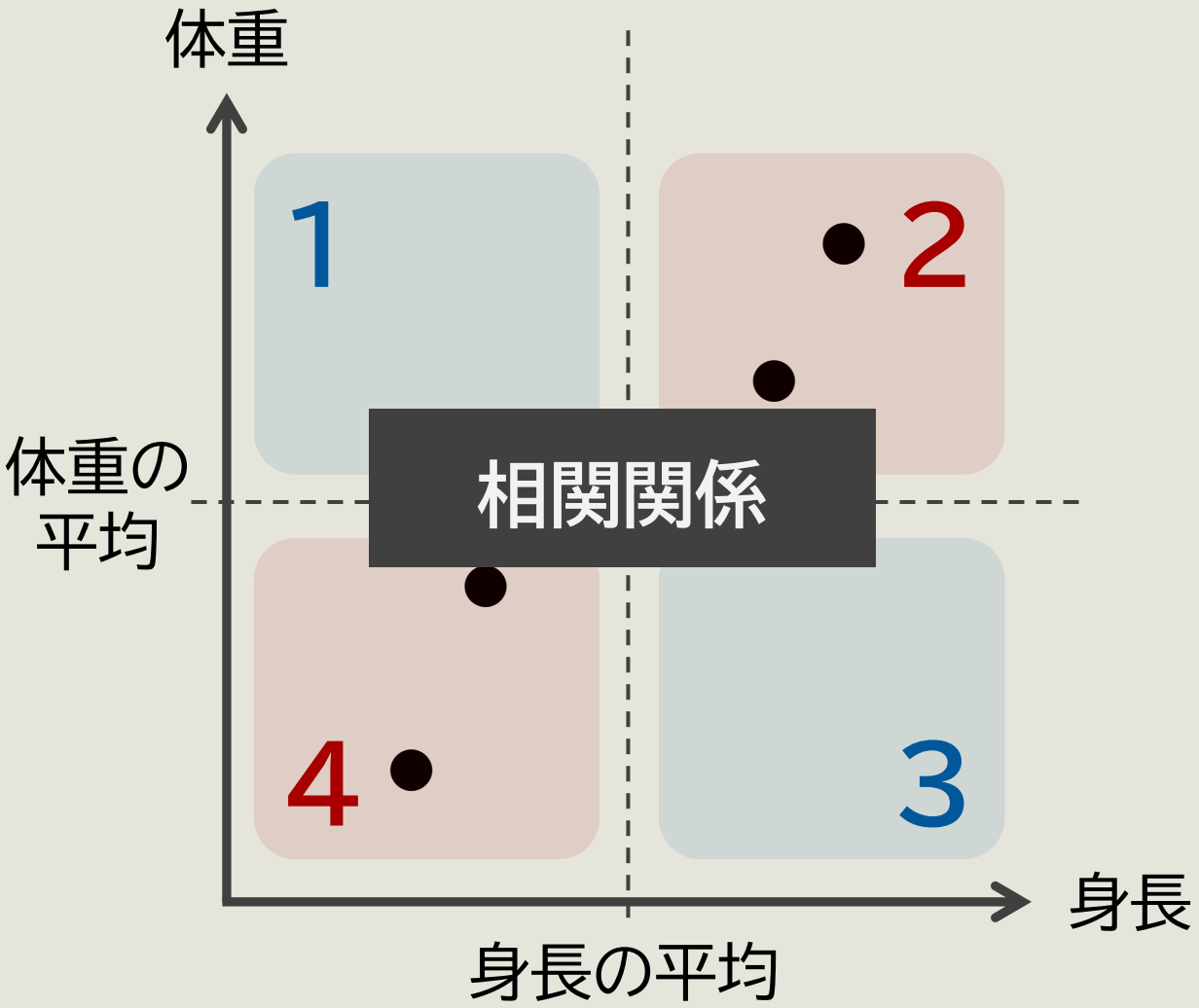
セクション1：相関分析

散布図は2変数のデータのばらつきをみる図





相関関係は因果関係の1条件



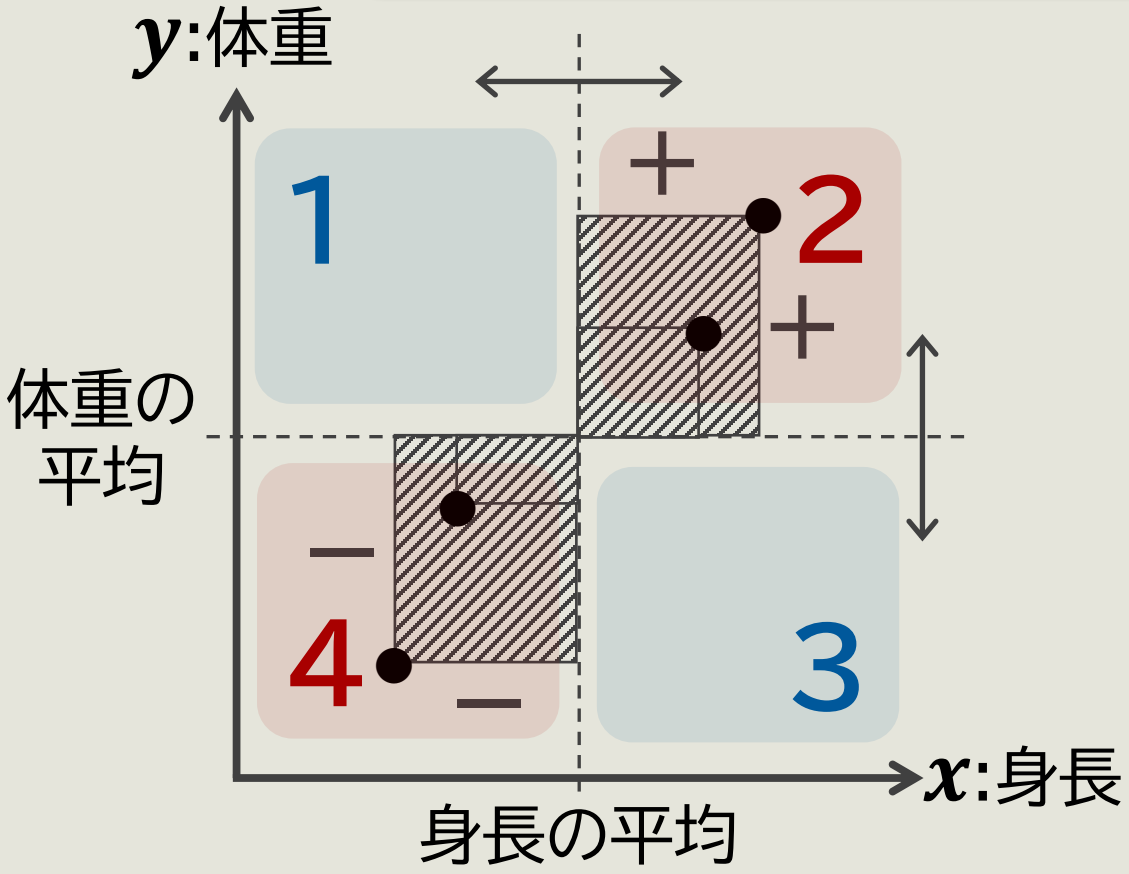
相関関係

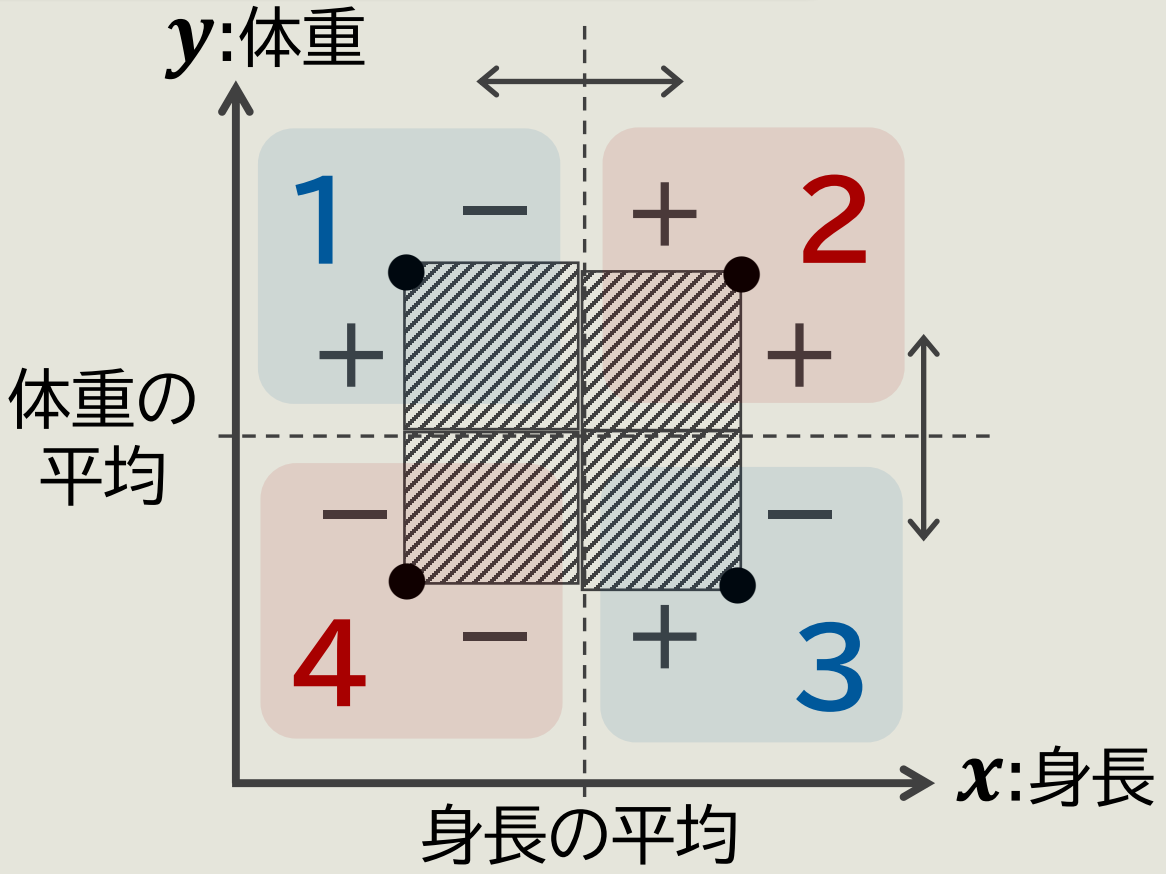
因果関係

- ☑ 相関関係
- ☑ 時間的先行(原因と結果)
- ☑ 第3因子によらない

相関関係を数値で表現したい…。そこで共分散！

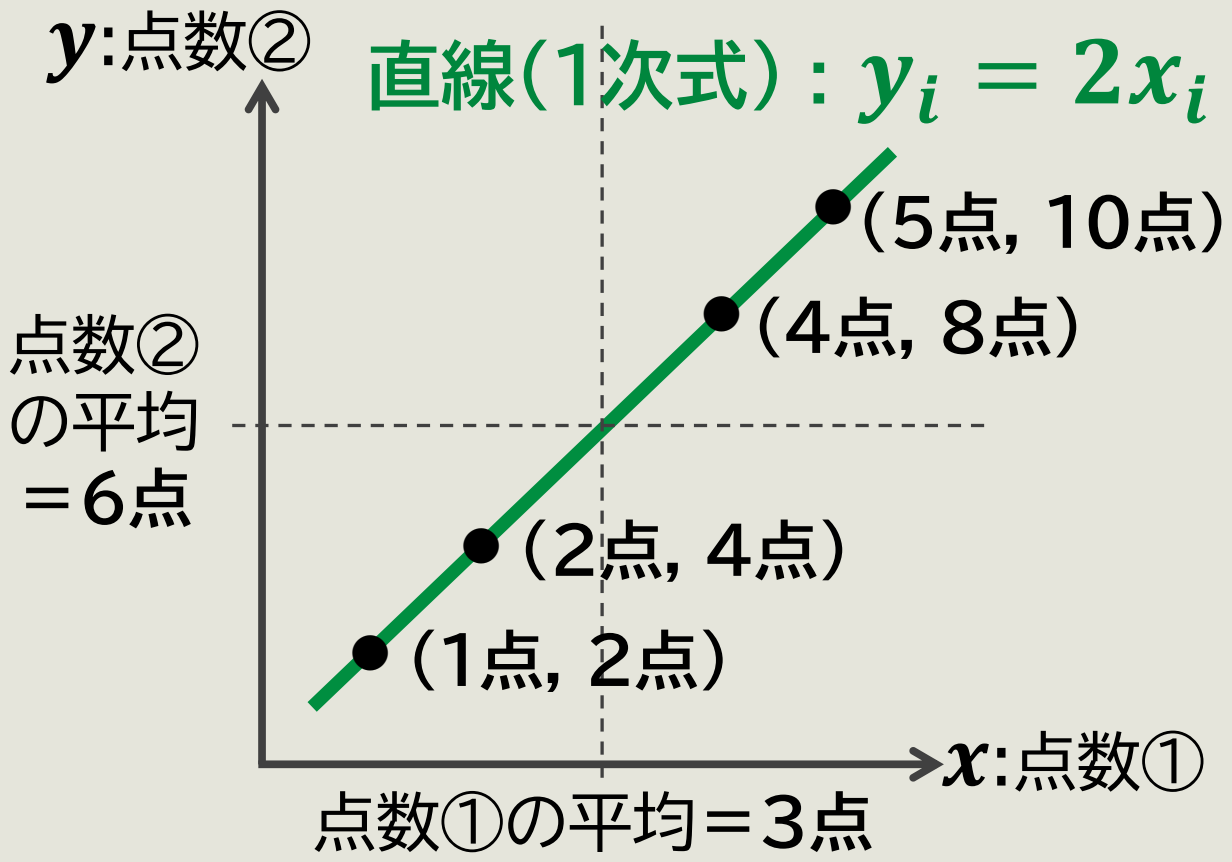
$$\text{共分散}：s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$





共分散 s_{xy} が最大/最小となるときとは？

完全な相関関係



共分散(完全な相関関係のとき)

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(2x_i - 2\bar{x})$$

$$= 2 \times \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$= 2{s_x}^2 = s_x \times 2s_x = s_x s_y$$

$$-s_x s_y \leq s_{xy} \leq s_x s_y$$

共分散 s_{xy} にモノサシを与える…。それが相関係数 r

$$-s_x s_y \leq s_{xy} \leq s_x s_y$$

モノサシ(分母)を
$s_x s_y$にそろえる

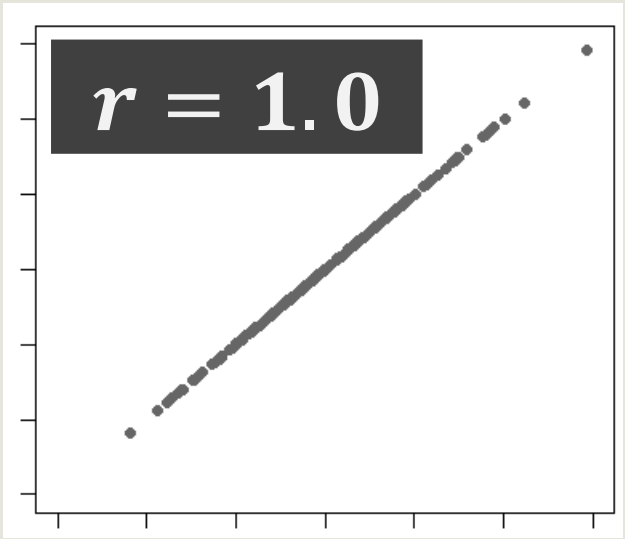
$$-100 \leq s_{xy} \leq 100$$
$$-30 \leq s_{yz} \leq 30$$
$$-5 \leq s_{xz} \leq 5$$
$$\vdots$$

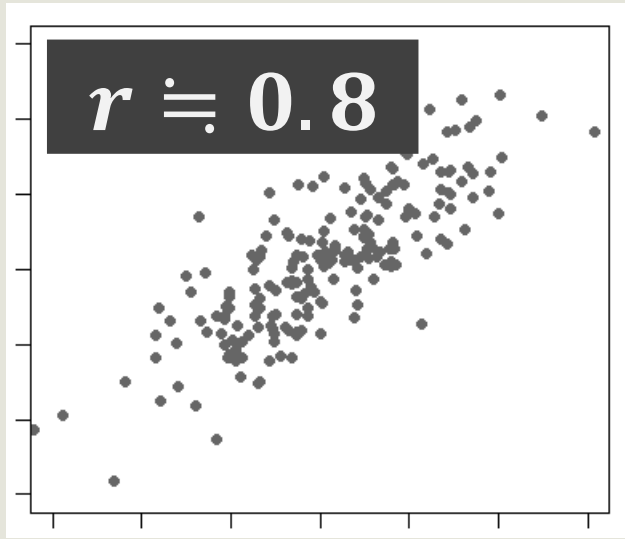
比較ができない…

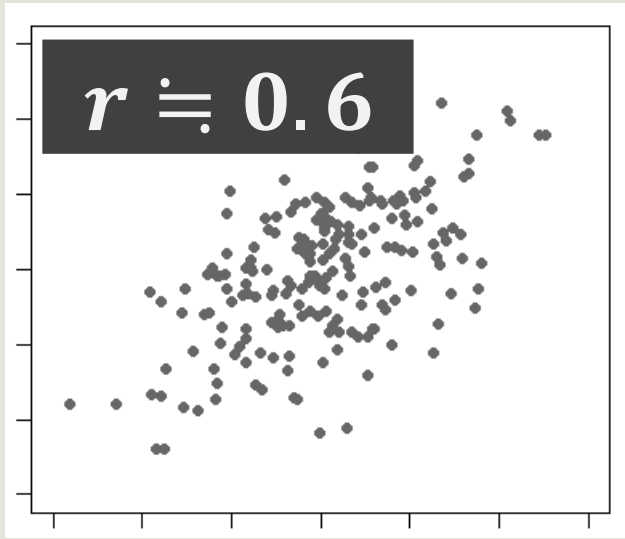
$$-1 \leq \frac{s_{xy}}{s_x s_y} \leq 1$$

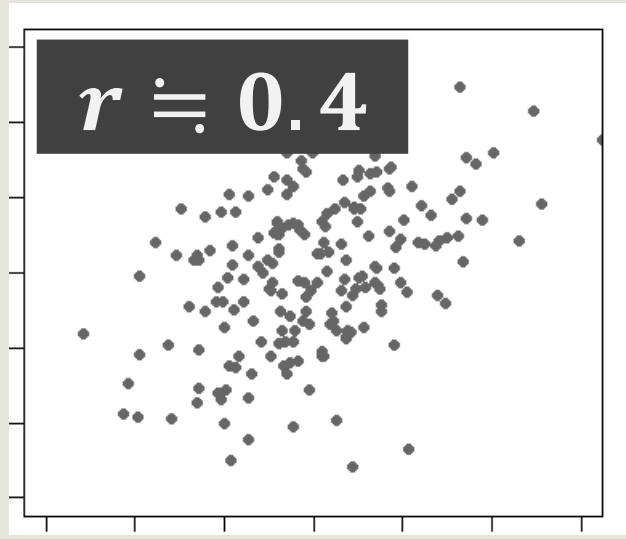
相関係数 r

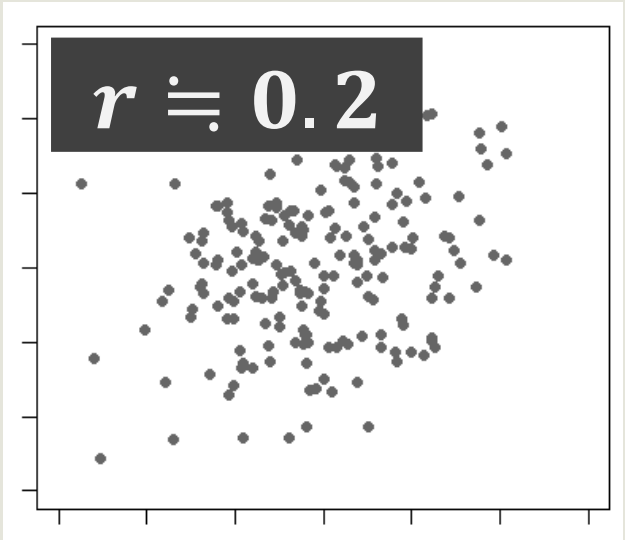
散布図から相関係数 r を(おおよそ)把握できる

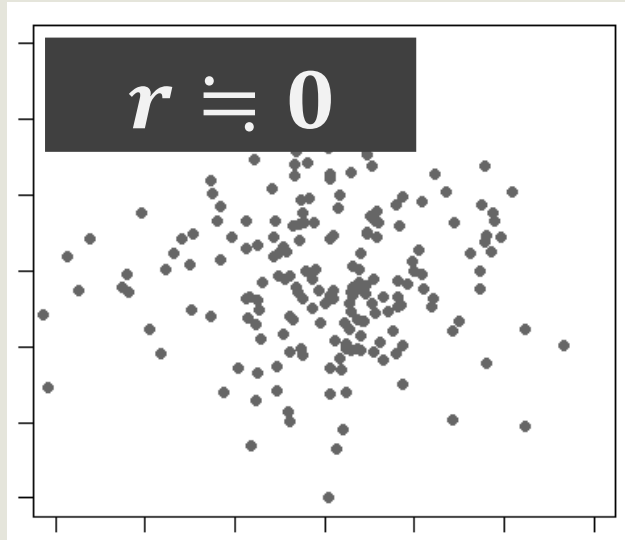


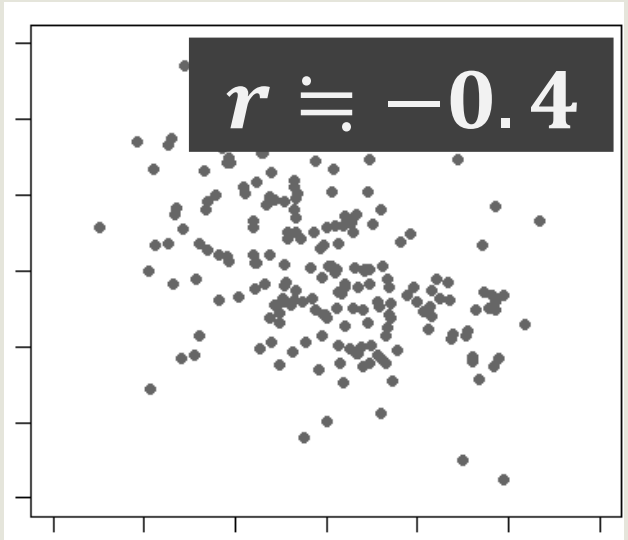


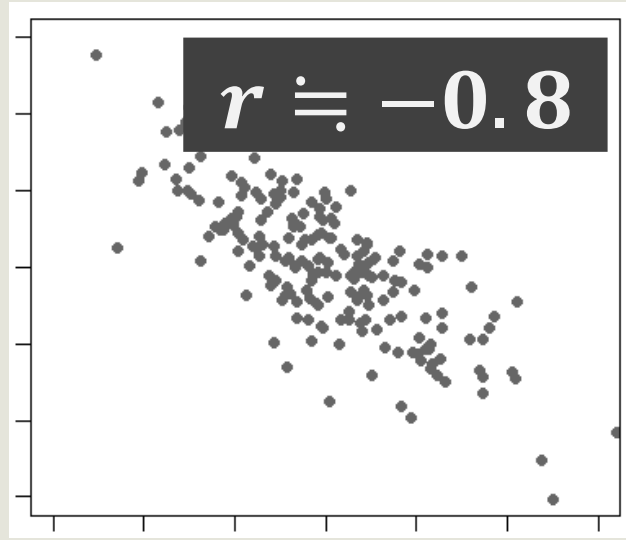






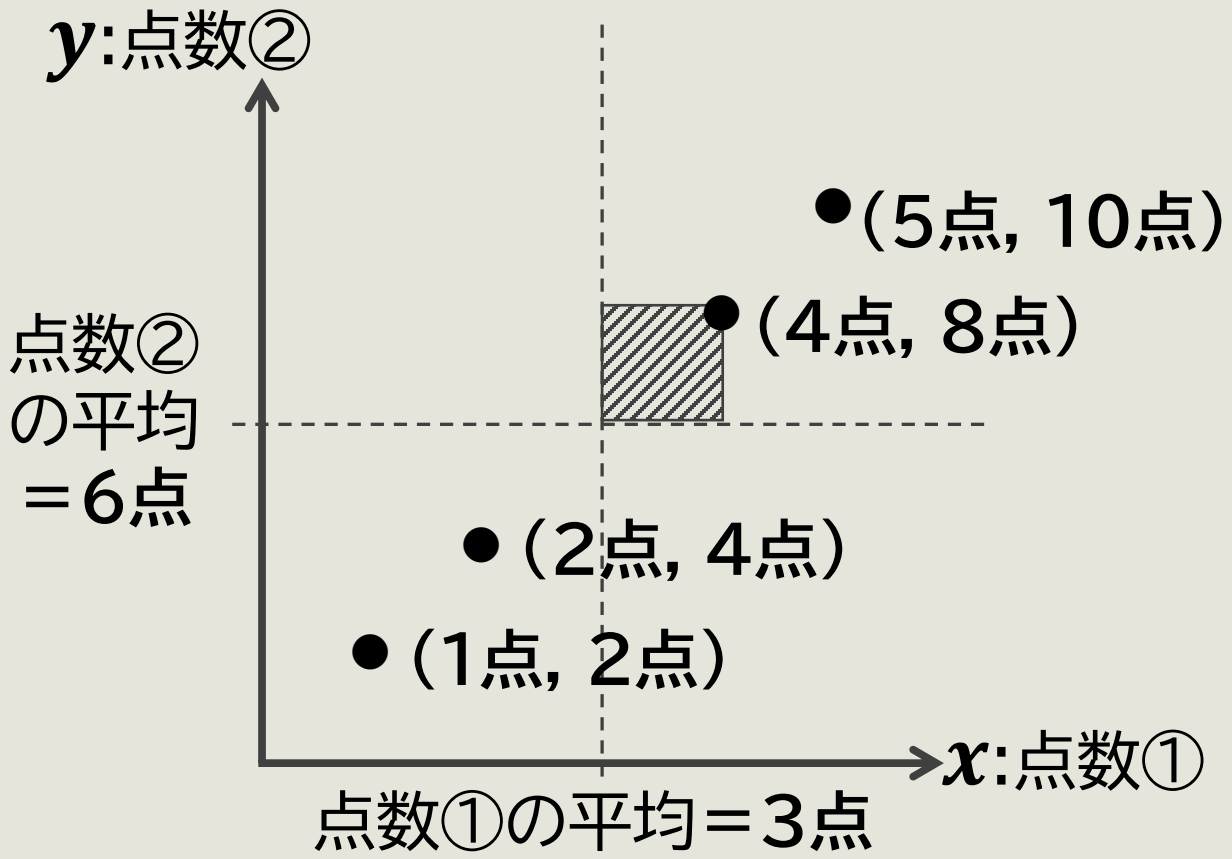


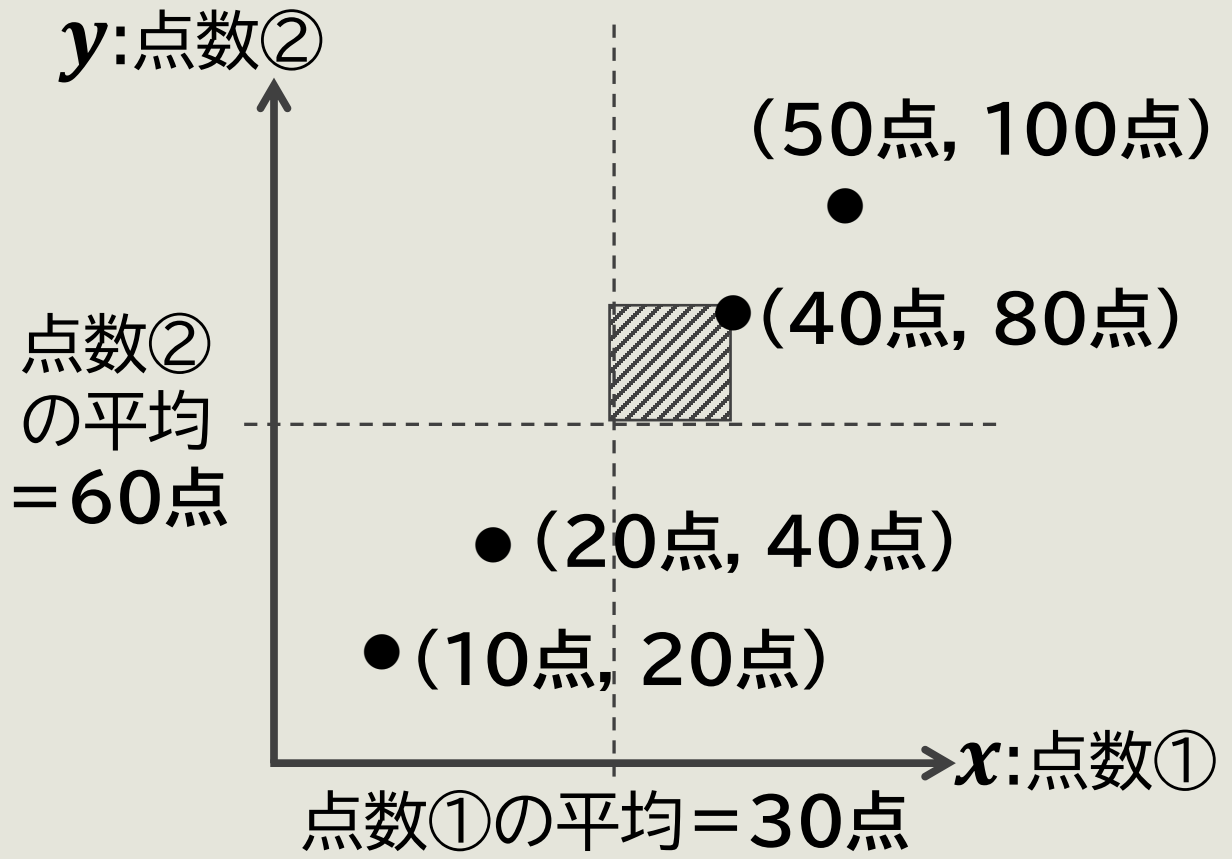




単位に依存する共分散 s_{xy}、依存しない相関係数 r

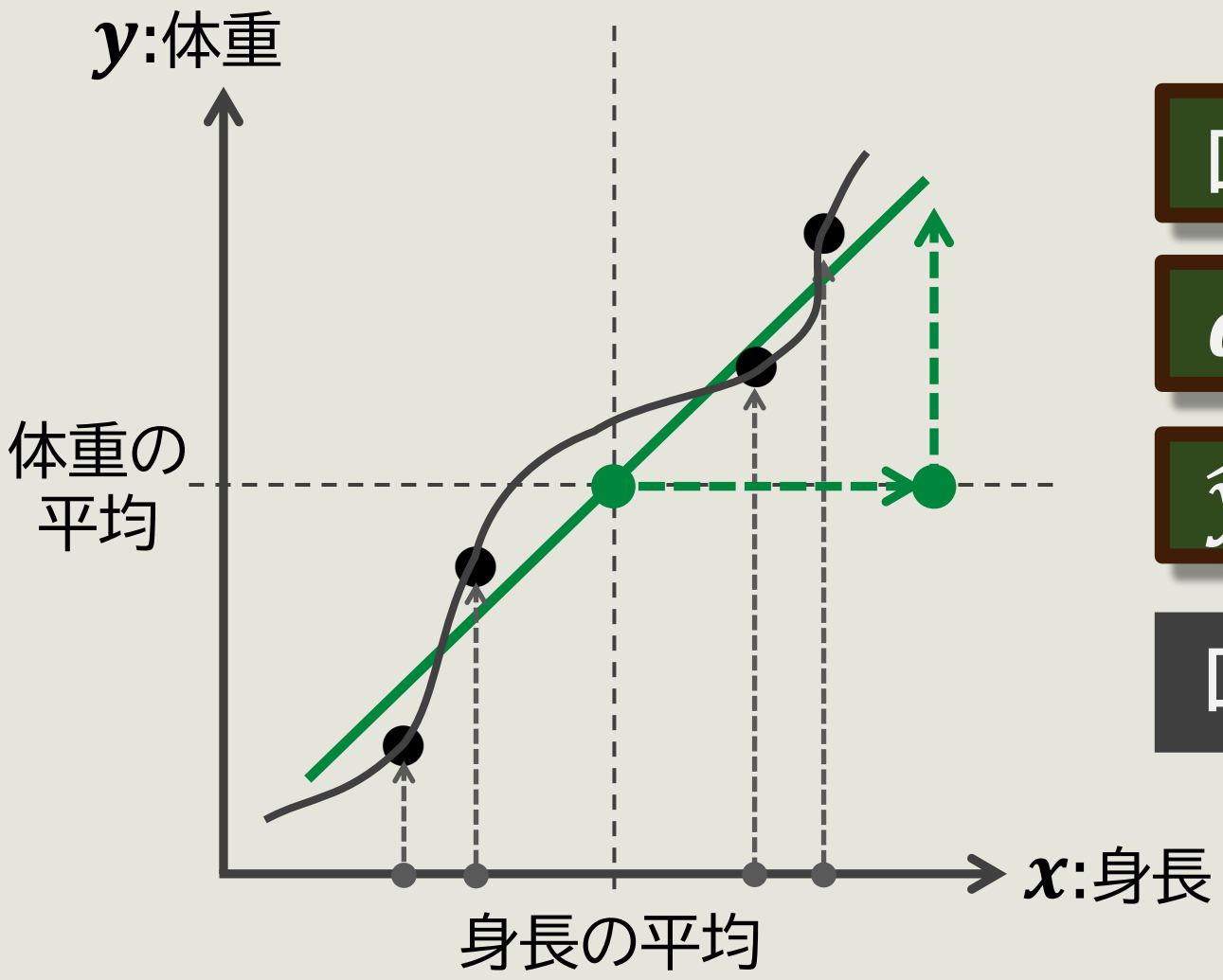
$$s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}),\ r = s_{xy}/s_x s_y$$





セクション2：単回帰分析

xの値を与えたときのyの平均(条件つき平均)を「直線」で考える



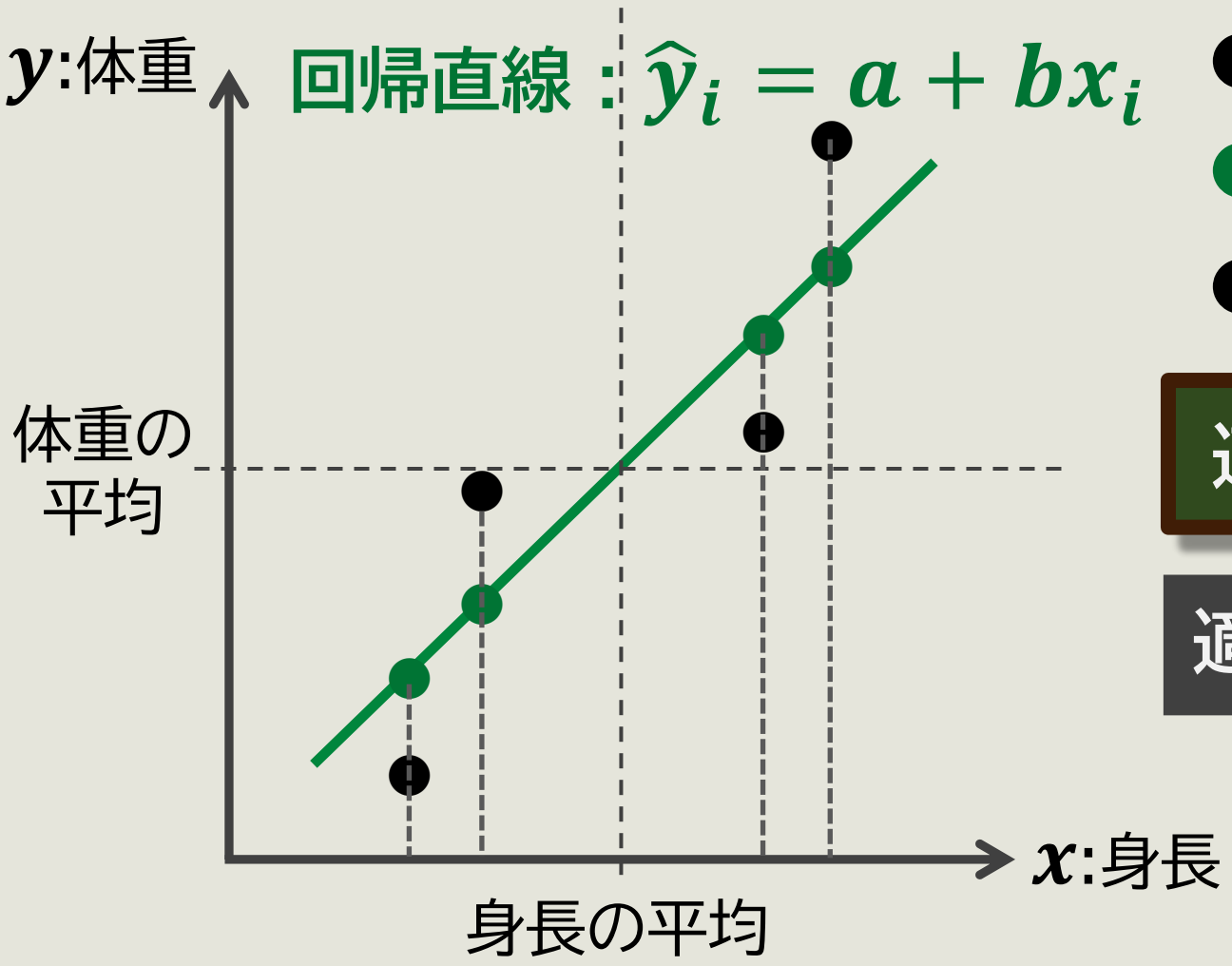
回帰直線：$\hat{y} = a + bx$

a：切片，　b：傾き(回帰係数)

$\hat{y}$：回帰直線による予測値

回帰係数bは相関関係の指標の1つ

切片 a と回帰係数 b をどう決めようか…



●：実際の値（実測値）y_i

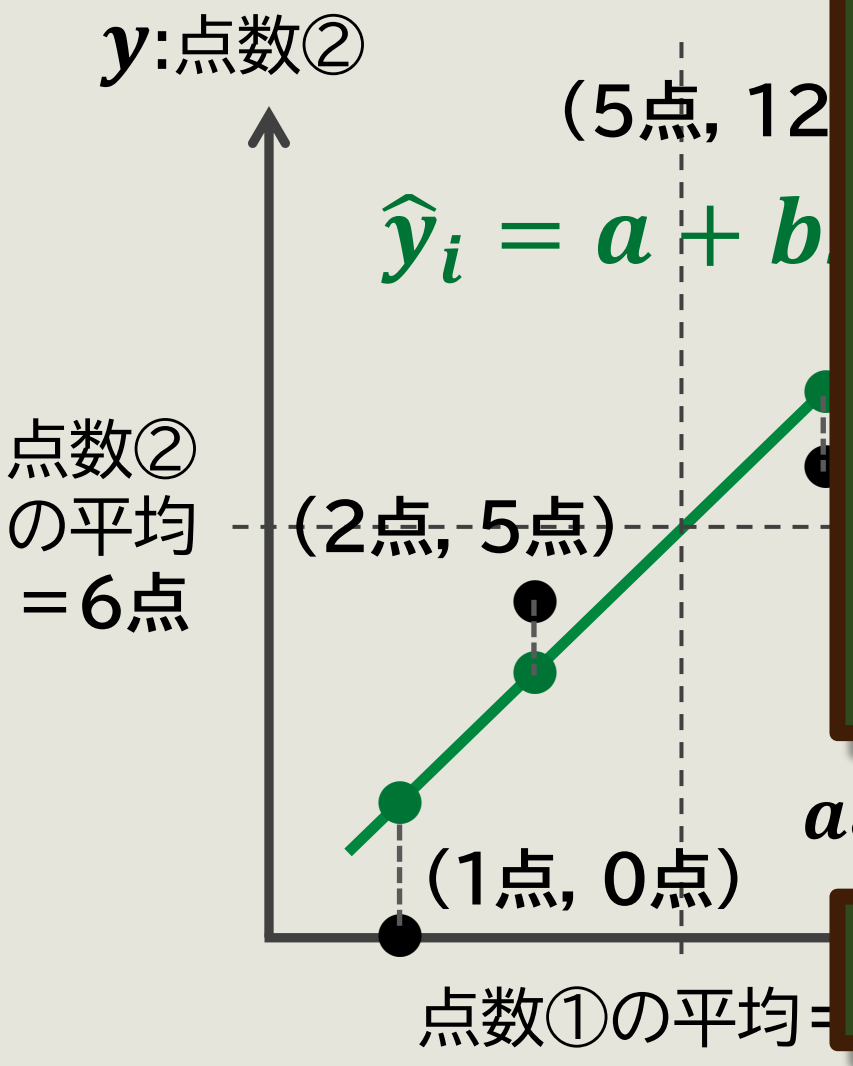
●：回帰直線による予測値 $\hat{y}_i$

●と●の差（距離）：残差 $y_i - \hat{y}_i$

適合の悪さ：$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

適合の悪さが最小となるa, bを探す

（偏）微分して切片 a と回帰係数 b についての連立方程式を解く



適合の悪さ：$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

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

$= (0 - a - 1b)^2 + (5 - a - 2b)^2$

$+ (7 - a - 4b)^2 + (12 - a - 5b)^2$

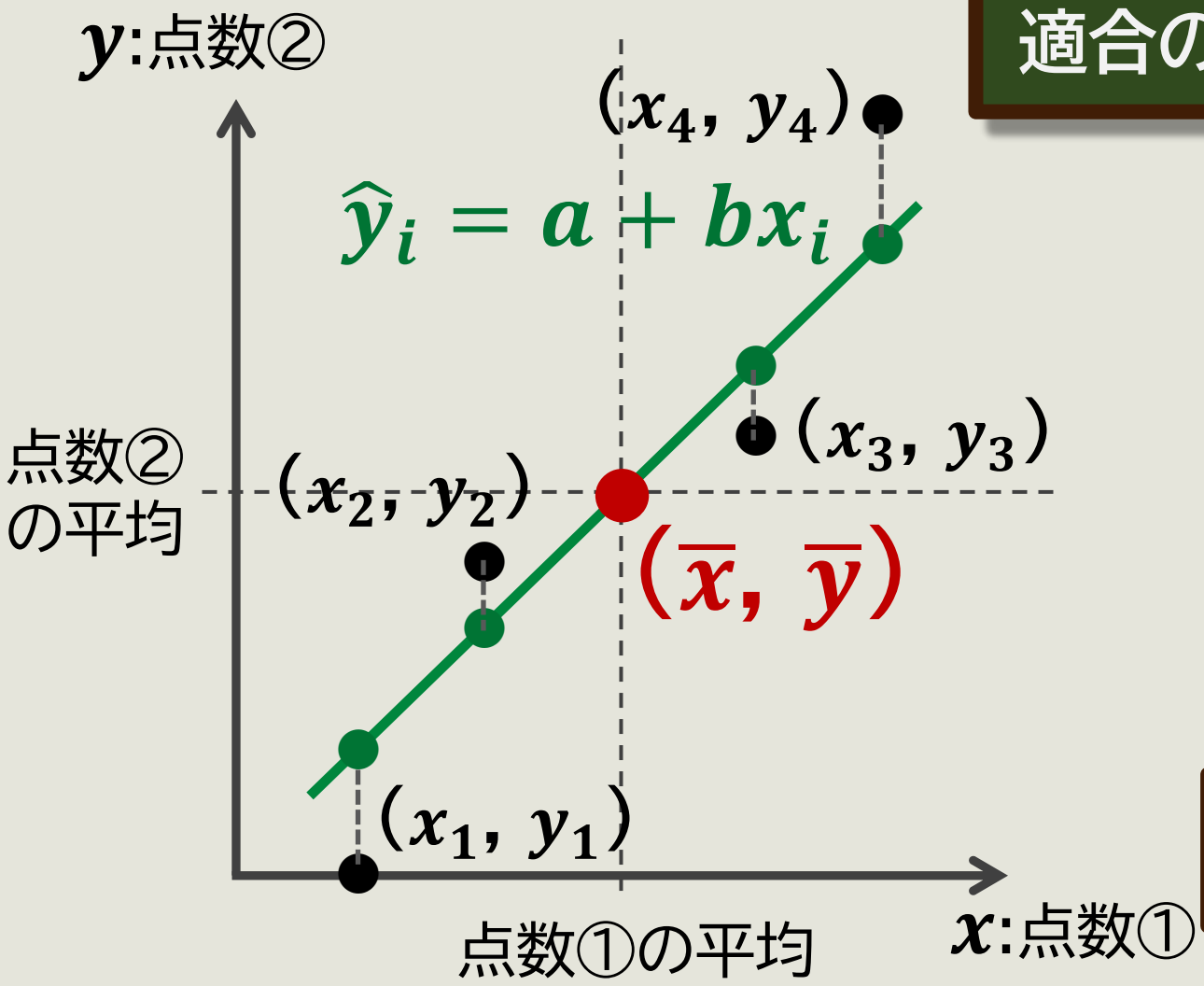
$= 4a^2 + 46b^2 + 24ab$

$-48a - 196b + 218$

aおよびbについて偏微分、結果を0とおき連立方程式を解く‥

$a = -1.8, b = 2.6$（最小二乗法による推定）

回帰直線は(xの平均, yの平均)を必ず通る



$$\text{適合の悪さ} = \sum_{i=1}^{N}(y_i - a - bx_i)^2$$

↓ 最小

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

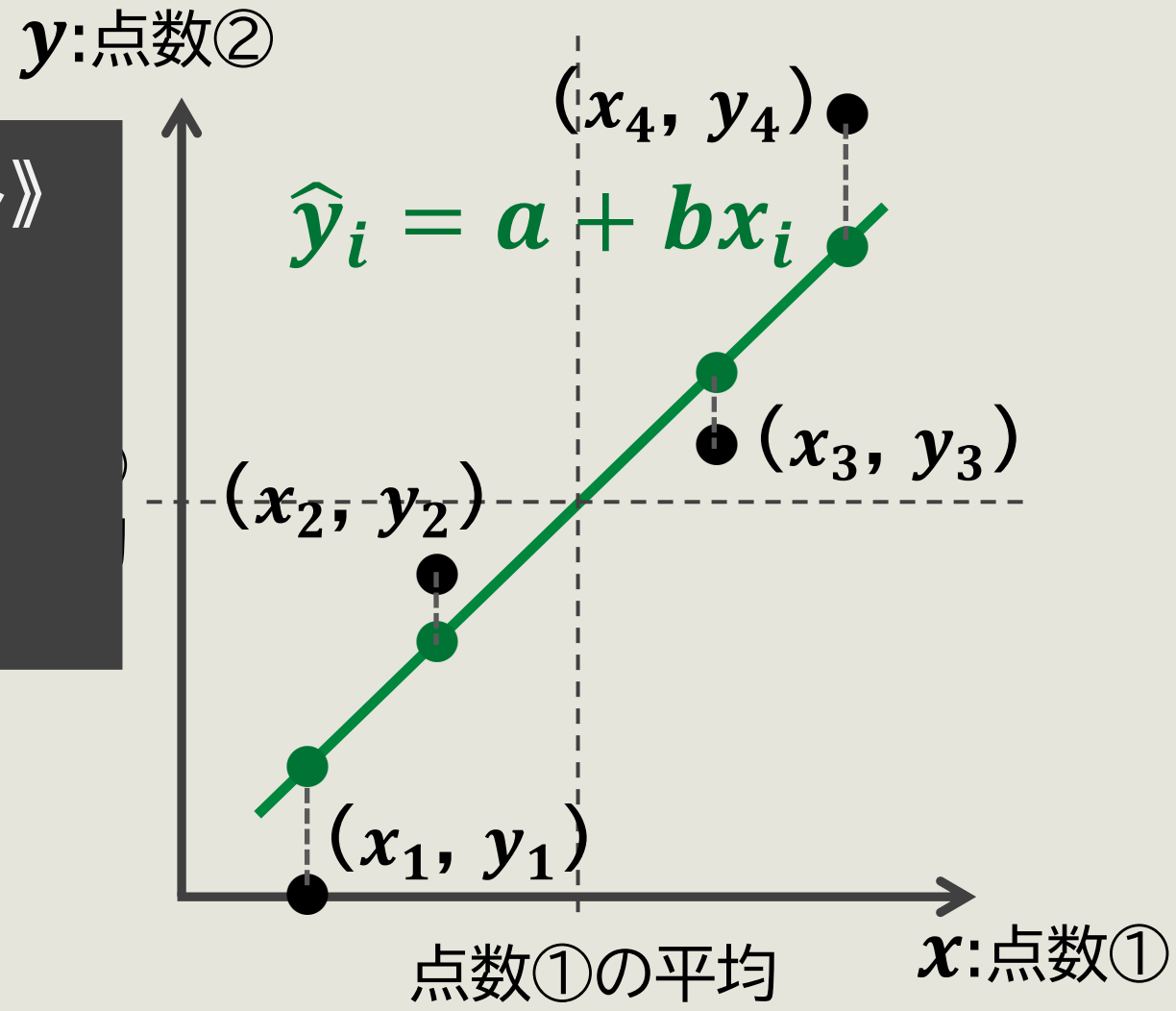
$$b = r \times \frac{s_y}{s_x}$$

↓ 回帰直線のaに代入

$$\hat{y}_i = \bar{y} + b(x_i - \bar{x})$$

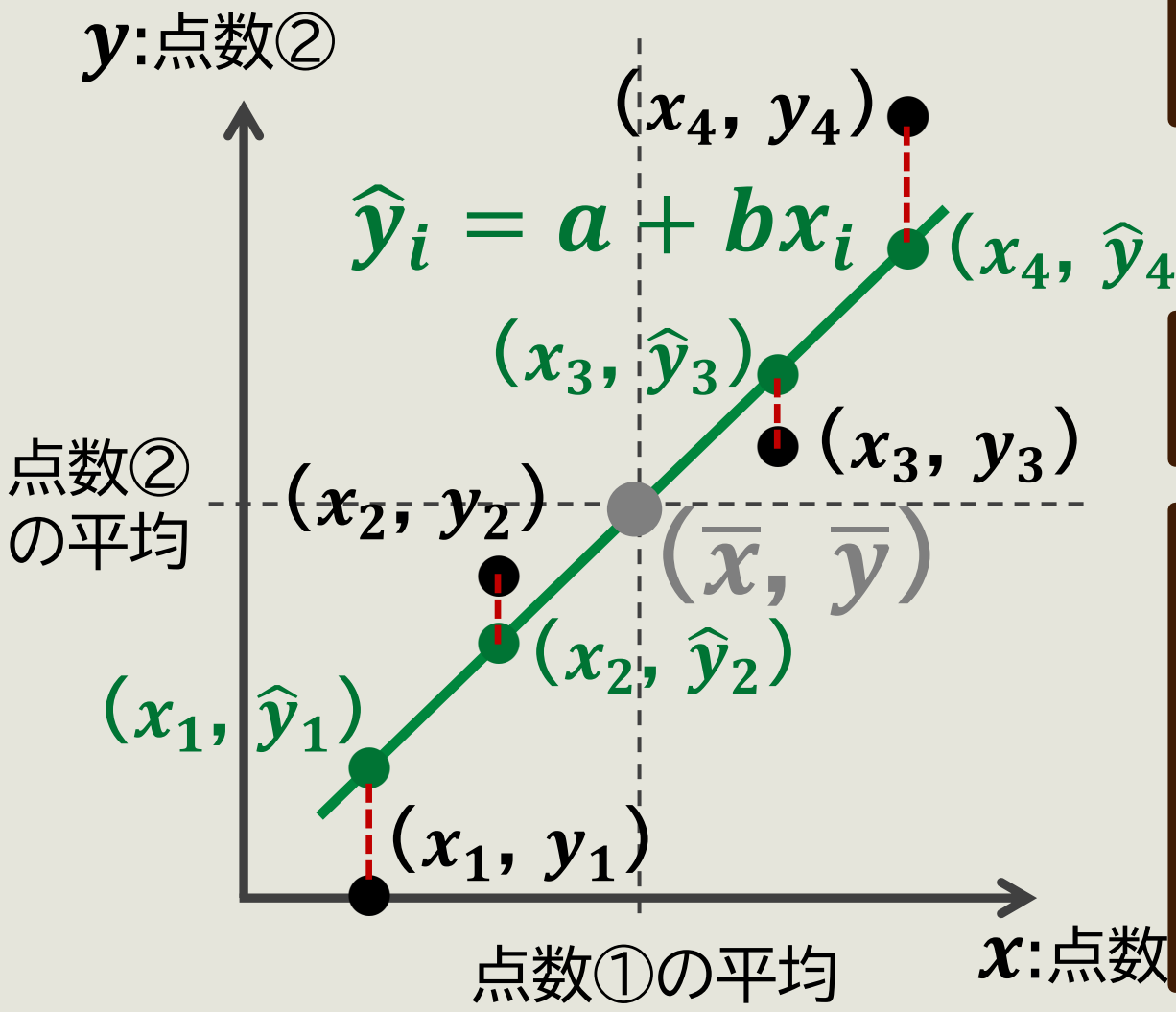
アウトプットとインプット

《アウトプット》
- 目的変数
- 被説明変数
- 従属変数
- ターゲット



《インプット》
- 説明変数
- 予測変数
- 独立変数
- 特徴量

実測値 y_i と予測値 $\hat{y}_i$ のズレを残差という



適合の悪さ：$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

→ 残差

残差：$e_i = y_i - \hat{y}_i$

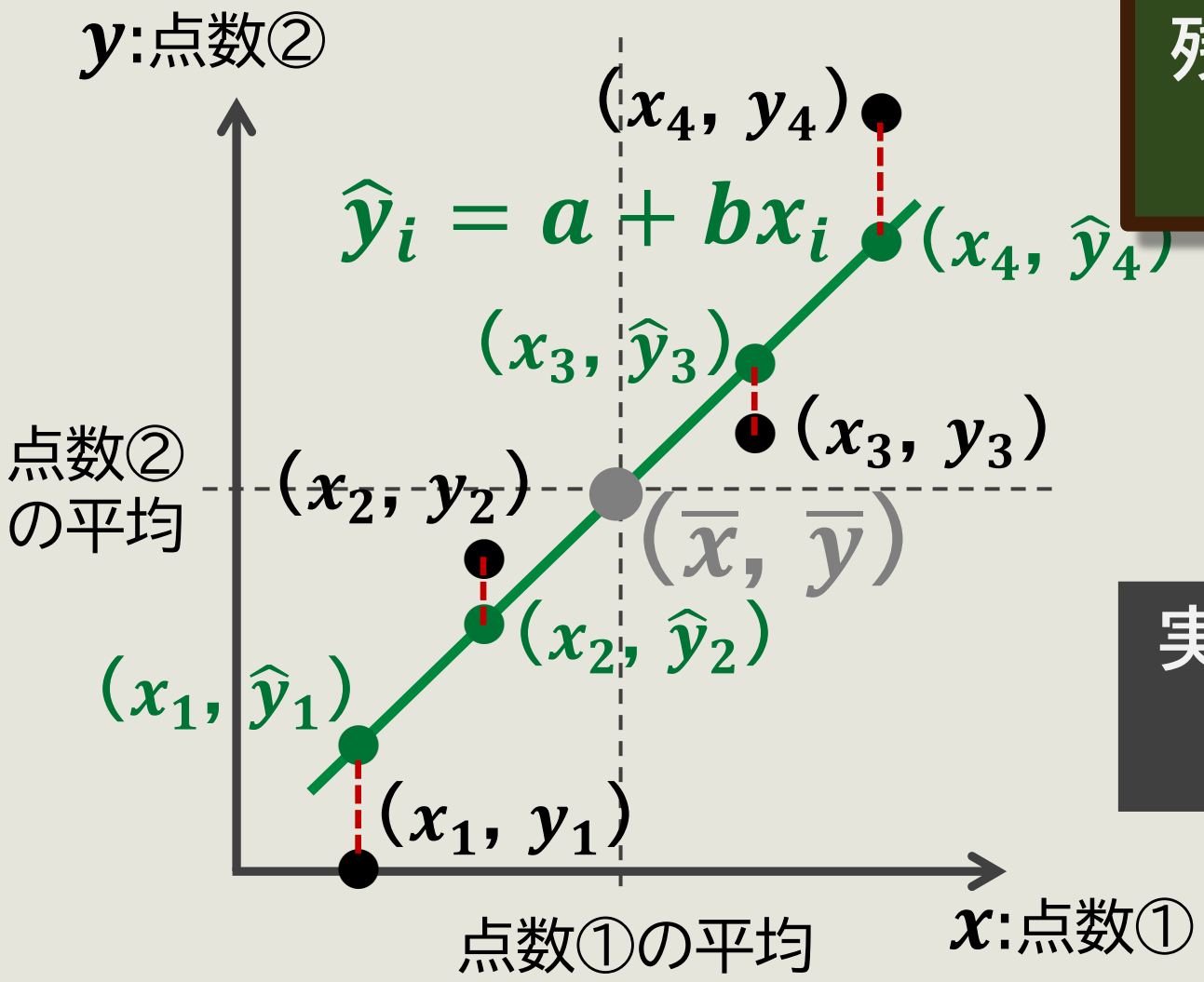
《残差の性質》

① 残差の平均：$\bar{e} = \bar{y} - \bar{\hat{y}} = 0$

② x_iとe_iは無相関：$r_{xe} = 0$

③ $\hat{y}_i$とe_iも無相関：$r_{\hat{y}e} = 0$

回帰直線は実測値 y を予測値 $\hat{y}$ と残差 e に分解するもの



残差 : $e_i = y_i - \hat{y}_i$

$\rightarrow y_i = \hat{y}_i + e_i$

→ 直交分解の式

実測値 y_i

予測値 $\hat{y}_i$

x_iで完全に説明できる部分

↕無相関

残差 e_i

x_iでは全く説明できない部分

残差 e は好ましくない邪魔なもの？

残差：$e_i = y_i - \hat{y}_i$

実際の価格

駅からの距離で説明できる価格

《残差の意味》
実際の価格のうち、駅からの距離では説明がつかない価格

実測値 y のばらつきを予測値 $\hat{y}$ がどれくらい説明できたかを考える

実測値：y_i

分解

予測値：$\hat{y}_i = a + bx_i$

↕無相関

残差：e_i

x_iでは全く説明できない部分

データのばらつき方に着目すると

y_iの分散：${s_y}^2$

分解

$\hat{y}_i$の分散：${s_{\hat{y}}}^2 = (b \times s_x)^2 = {s_y}^2 r^2$

y_iの分散のうちx_iで完全に説明できる分散

e_iの分散：${s_e}^2 = {s_y}^2 - {s_{\hat{y}}}^2 = {s_y}^2(1 - r^2)$

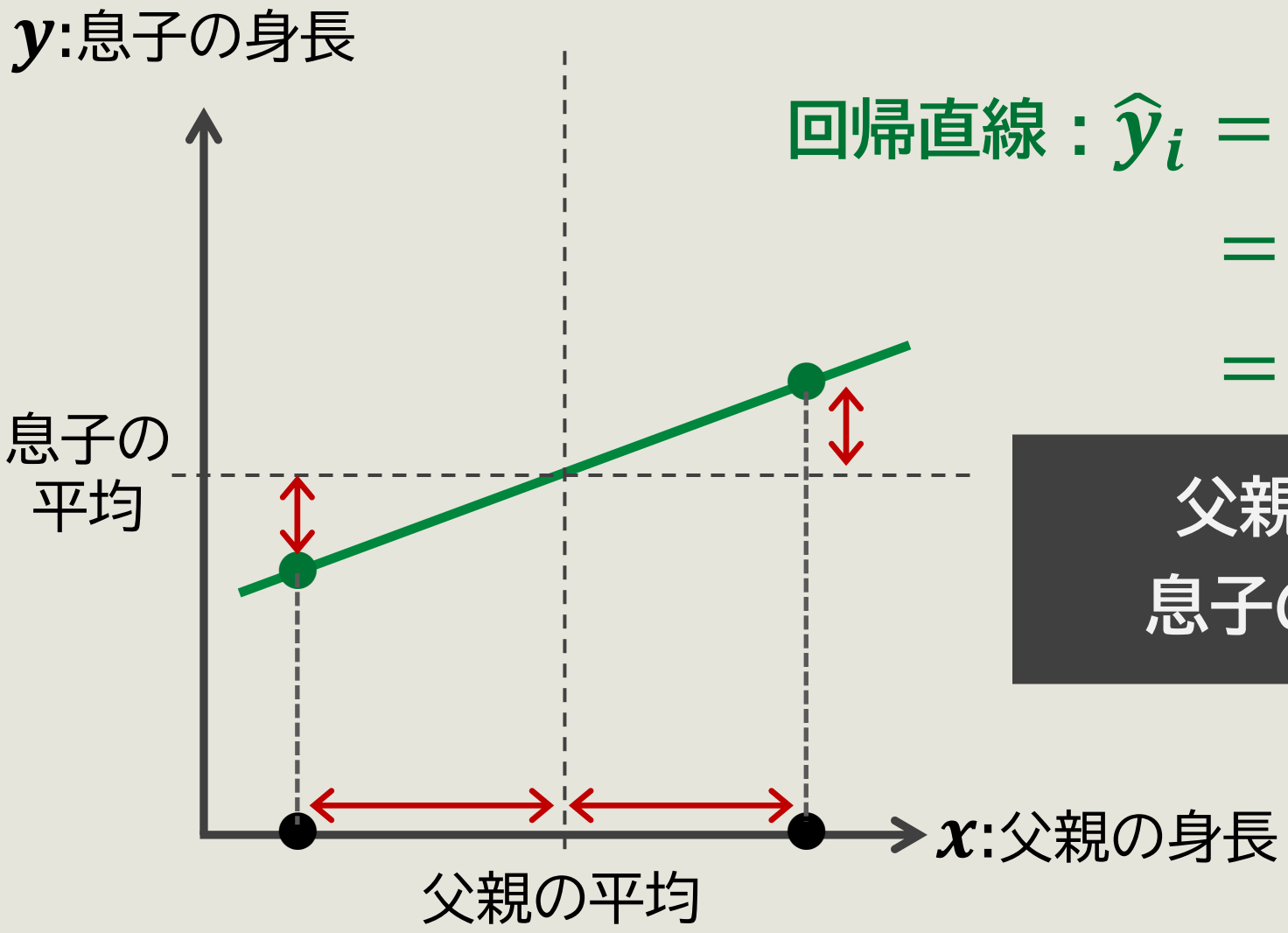
y_iの分散のうちx_iでは全く説明できない分散

分散説明率(決定係数)：r^2

yの分散：${s_y}^2$

$\times r^2$	$\times (1 - r^2)$
xで完全に説明できる部分	xでは全く説明できない部分

「回帰」とは平均(平凡)への「回帰」



回帰直線：$\hat{y}_i = a + bx_i$

$= \bar{y} + b(x_i - \bar{x})$

$= 170 + 0.7(x_i - 170)$

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

父親の身長(x)が非平凡でも
息子の身長(y)は平凡へ近づく

$b = r \times \frac{s_y}{s_x}$

セクション3：単回帰分析の視覚的理解

「偏差」変数をベクトルで表現すると都合が良い！

$\vec{x}$: xベクトル(向きと大きさを持つ)

$$\vec{x} = (168, 176)$$

1人目　2人目

$\vec{x}$ の大きさ: $\|\vec{x}\|$

$$= \sqrt{168^2 + 176^2}$$

$$\vec{x} = (168 - \bar{x}, 176 - \bar{x})$$
$$= (168 - 172, 176 - 172)$$
$$= (-4, 4)$$　平均からの偏差ベクトル

$\vec{x}$ の大きさ: $\|\vec{x}\| = \sqrt{(-4)^2 + 4^2}$

xの分散：$s_x{}^2 = \|\vec{x}\|^2/n$

xの標準偏差：$s_x = \|\vec{x}\|/\sqrt{n}$

ベクトルの内積を使うとさらに都合が良い！

1人目　2人目

身長：$\vec{x} = (168, 176)$

体重：$\vec{y} = (\ 70,\ \ 74\ )$

身長の偏差：$\vec{x} = (168 - \overline{x}, 176 - \overline{x})$

体重の偏差：$\vec{y} = (70 - \overline{y}, 74 - \overline{y})$

《偏差ベクトルの内積》

1人目　2人目

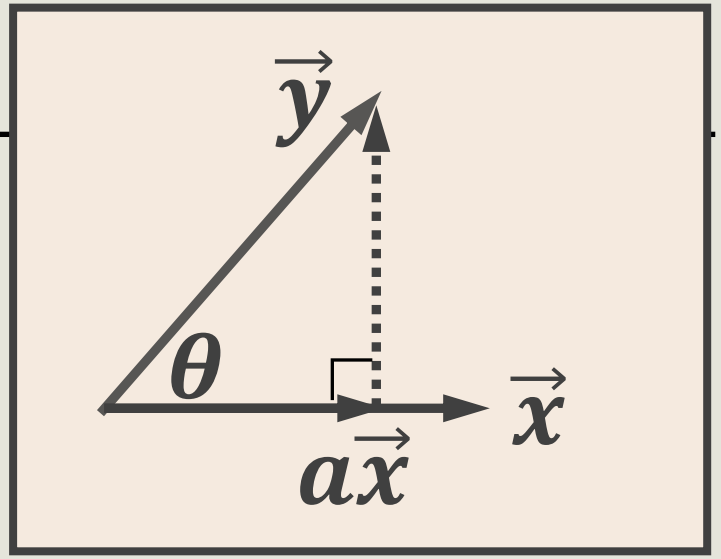
$$\vec{x} \cdot \vec{y} = (168 - \overline{x})(70 - \overline{y}) + (176 - \overline{x})(74 - \overline{y})$$

共分散：$s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$

偏差ベクトルの内積：$\vec{x} \cdot \vec{y} = n \times s_{xy}$

相関係数 $\cos\theta$ ①

相関係数 $=\cos\theta$



(偏差)ベクトルの内積 : $\vec{x}\cdot\vec{y}=n\times s_{xy}$

《(偏差)ベクトルの内積の(もうひとつの)定義》

$$\vec{x}\cdot\vec{y}=\|\vec{x}\|\times\|\vec{y}\|\times cos\,\theta$$

xの標準偏差 : $s_x=\|\vec{x}\|/\sqrt{n}\rightarrow\|\vec{x}\|=\sqrt{n}s_x$

$$=\sqrt{n}s_x\times\sqrt{n}s_y\times cos\,\theta=ns_xs_y\,cos\,\theta=ns_{xy}$$

$cos\,\theta=s_{xy}/s_xs_y=r$: 相関係数

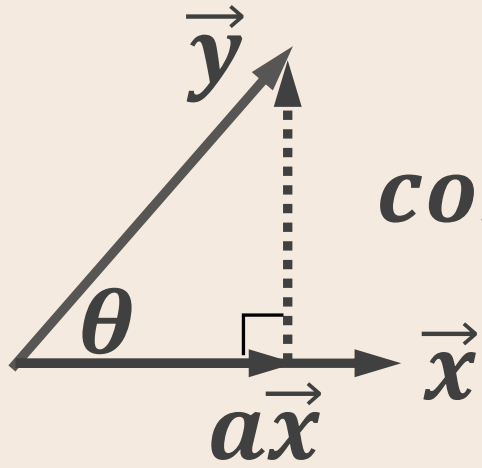
$$cos\,\theta=\frac{a\|\vec{x}\|}{\|\vec{y}\|}$$

相関係数 $\cos\theta$ ②

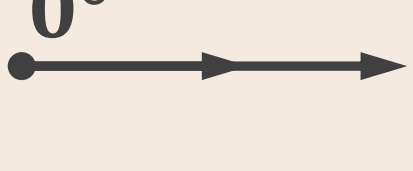
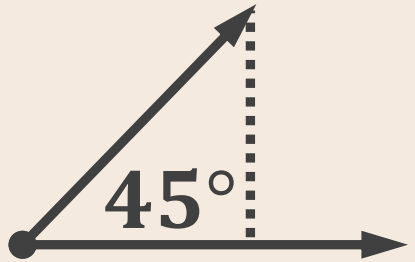
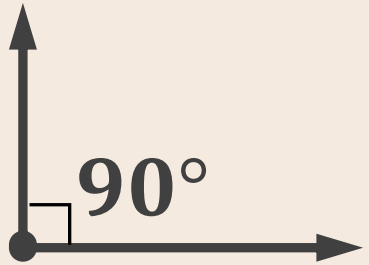
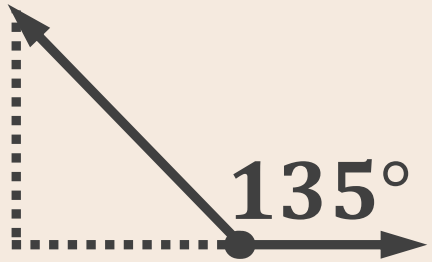
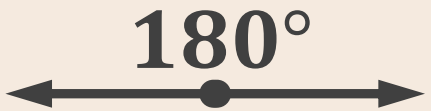
相関係数 $=$ co

$\cos\theta = s_{xy}/s_x s_y = r$：相関係数

$-1 \leq \cos\theta \leq 1$



$$\cos\theta = \frac{a\|\vec{x}\|}{\|\vec{y}\|}$$

$\cos 0° = 1$	$\cos 45° \fallingdotseq 0.7$	$\cos 90° = 0$	$\cos 135° \fallingdotseq -0.7$	$\cos 180° = -1$
				

2つのベクトルが直角のときに無相関となる。
2つのベクトルの向きが同じ（または正反対）のときに完全相関となる。

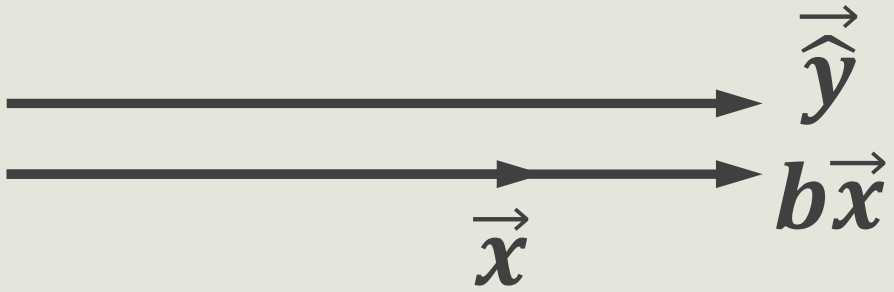
予測値ベクトル $\vec{\hat{y}}$ は、説明変数ベクトル $\vec{x}$ と同じ向き

予測値の偏差ベクトル：$\vec{\hat{y}} = (\hat{y}_1 - \overline{\hat{y}}, \hat{y}_2 - \overline{\hat{y}}, \dots)$

$$\hat{y}_i = \overline{y} + b(x_i - \overline{x}) \rightarrow \hat{y}_i - \overline{y} = b(x_i - \overline{x})$$

$$\vec{\hat{y}} = b \times (x_1 - \overline{x}, x_2 - \overline{x}, \dots) = b \times \vec{x}$$

$\vec{x}$：説明変数xの偏差ベクトル



予測値の偏差ベクトル$\vec{\hat{y}}$と
説明変数の偏差ベクトル$\vec{x}$は
同じ向き（or 正反対の向き）

$\vec{\hat{y}}$と$\vec{x}$は
完全相関

残差ベクトル $\vec{e}$ は回帰係数 b の大きさでさまざまに変動

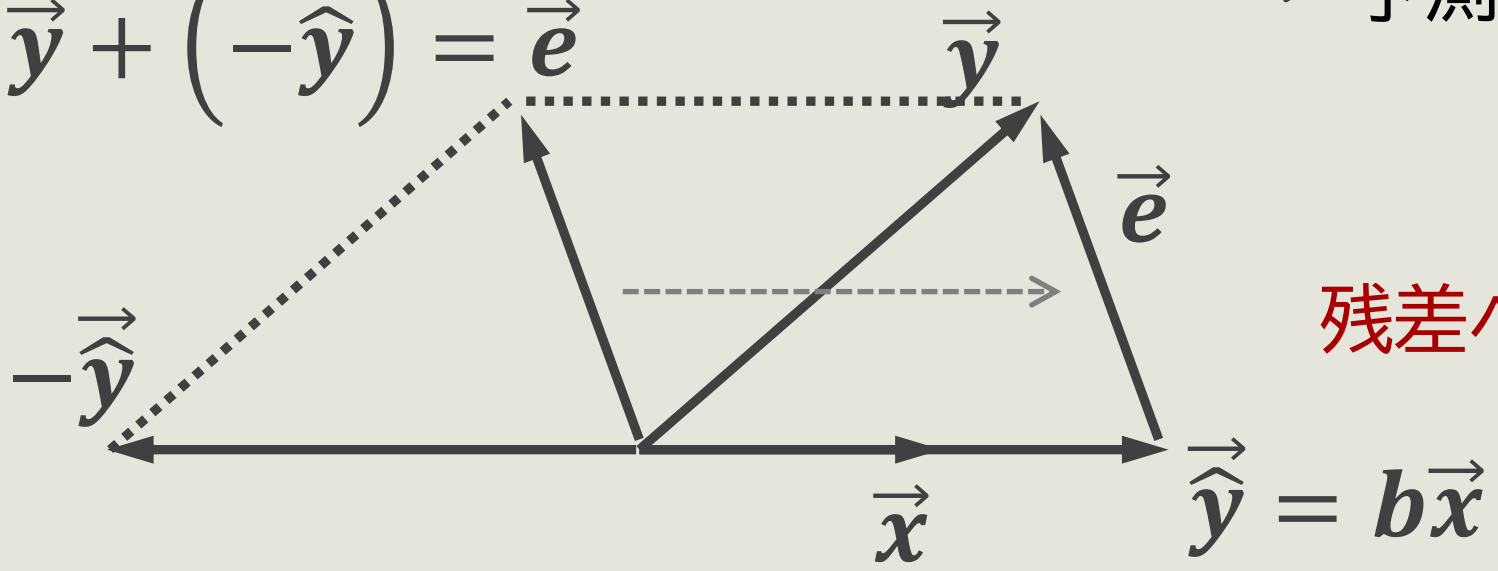
残差ベクトル：$\vec{e} = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, \cdots) = \vec{y} - \vec{\hat{y}}$

$(y_1 - \bar{y}) - (\hat{y}_1 - \bar{y})$

実測値の偏差ベクトル：$\vec{y}$

予測値の偏差ベクトル：$\vec{\hat{y}}$

$\vec{y} + \left(-\vec{\hat{y}}\right) = \vec{e}$



回帰係数$\boldsymbol{b}$によって

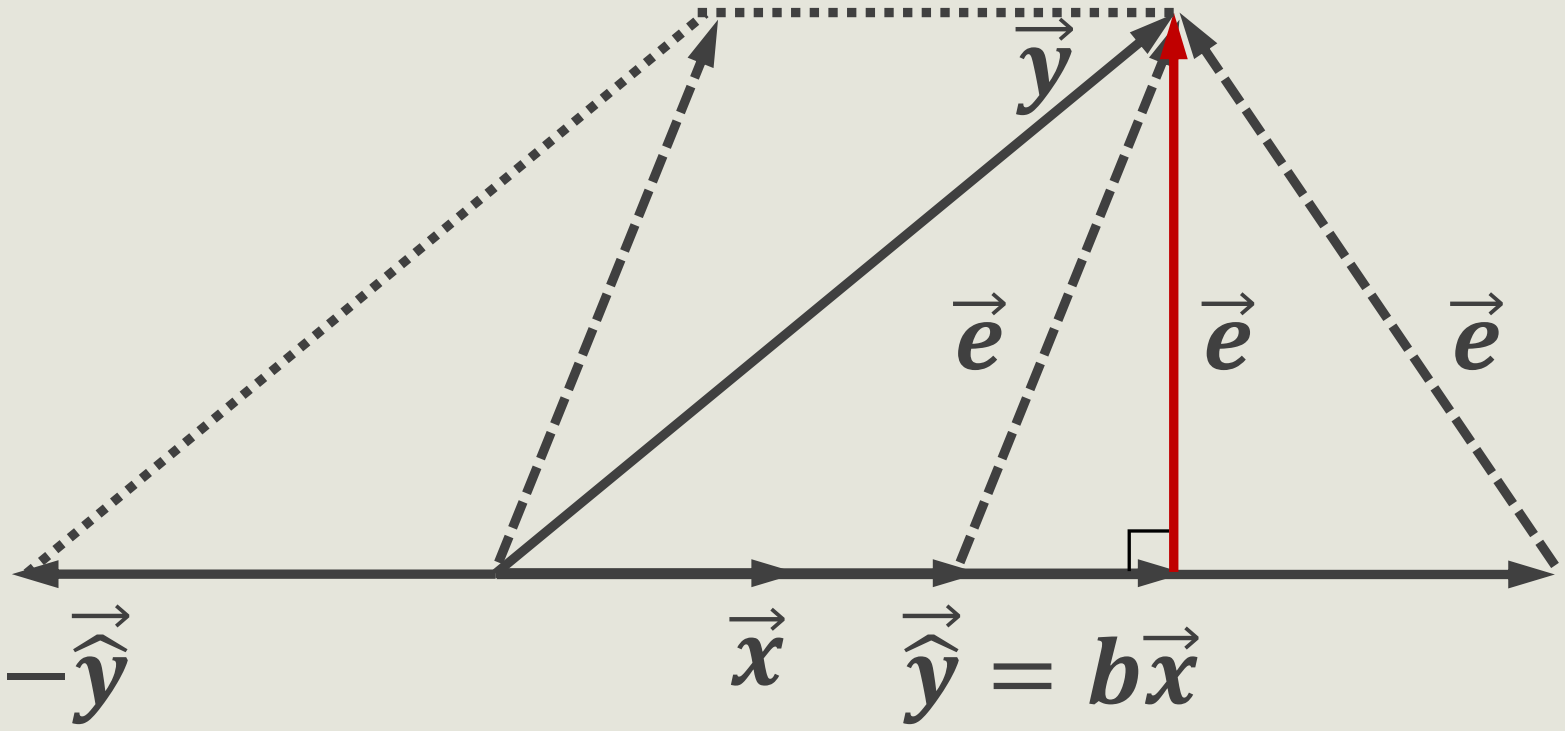
残差ベクトル$\vec{\boldsymbol{e}}$は変わっていく…

最小二乗法は残差ベクトル $\vec{e}$ を最短にする

残差ベクトル：$\vec{e} = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, \cdots) = \vec{y} - \hat{\vec{y}}$

適合の悪さ：$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \|\vec{e}\|^2$

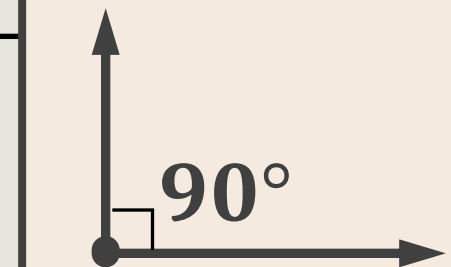
最小二乗法は残差ベクトル$\vec{e}$の大きさを最小にする…

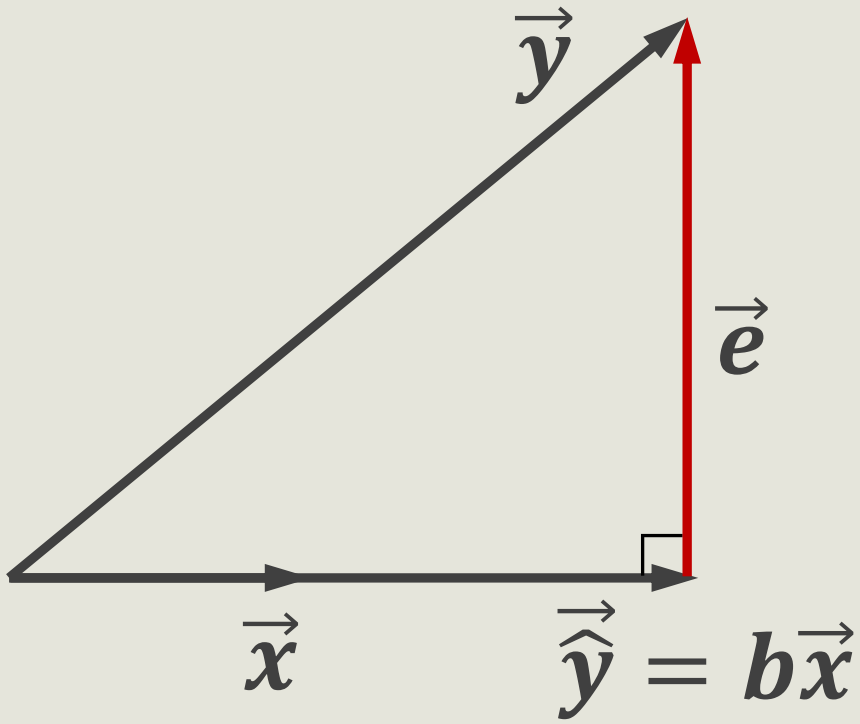


残差ベクトル$\vec{e}$の大きさが最小となるとき、
$\hat{\vec{y}}$と$\vec{e}$は直角（直交する）

予測値ベクトル $\vec{\hat{y}}$ と残差ベクトル $\vec{e}$ が「直交」す

$cos90° = 0$





残差ベクトル$\vec{e}$の大きさが
最小となるとき、
$\vec{\hat{y}}$と$\vec{e}$は直角になる（直交する）

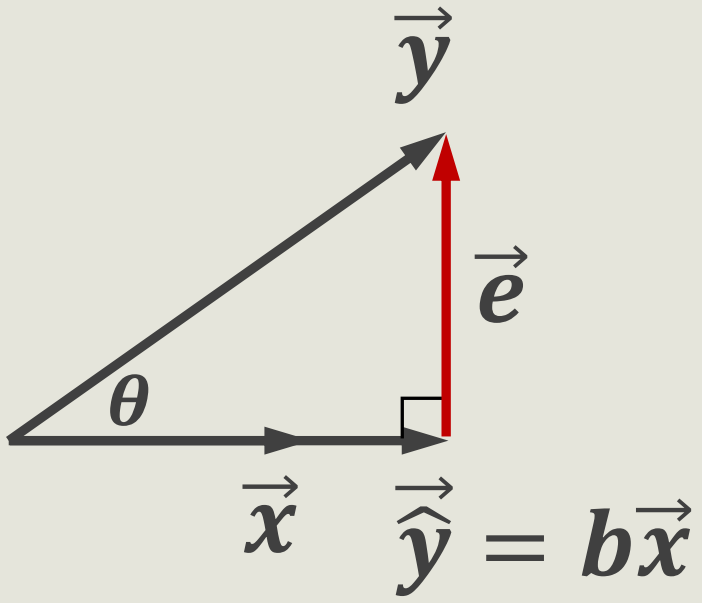
$$cos\,\theta = s_{xy}/s_x s_y = r : \text{相関係数}$$

残差：$e_i = y_i - \hat{y}_i$
$\rightarrow y_i = \hat{y}_i + e_i$

予測値$\hat{y}$と残差eの相関係数
：$r = cos\,90° = 0$

予測値$\hat{y}$と残差eは無相関（直交）の関係

分散説明率は $cos^2\,\theta$ 、つまり相関係数 r の2乗



$$\cos\theta = \frac{\left\|\hat{\vec{y}}\right\|}{\|\vec{y}\|}$$

$$\|\vec{y}\|^2 = \left\|\hat{\vec{y}}\right\|^2 + \|\vec{e}\|^2$$

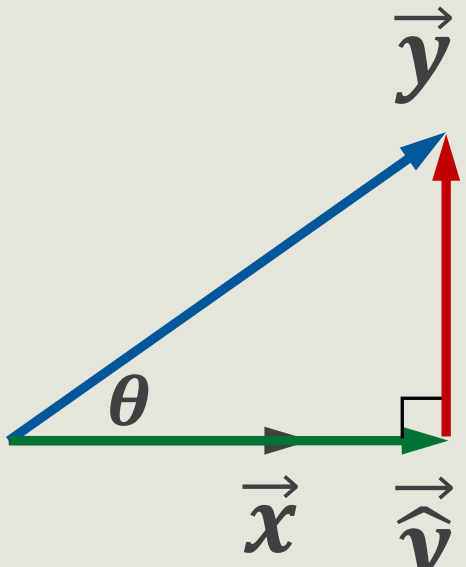
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

SS：平方和（*Sum of Squares*）

$$SS_y = SS_{\hat{y}} + SS_e$$

分散説明率：$SS_{\hat{y}}/SS_y = cos^2\,\theta = r^2$

分散説明率は $cos^2\theta$ 、つまり相関係数 r の2乗



$$\cos\theta = \frac{\|\hat{\vec{y}}\|}{\|\vec{y}\|}$$

$$\|\vec{y}\|^2 = \|\hat{\vec{y}}\|^2 + \|\vec{e}\|^2$$

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

SS：平方和（*Sum of Squares*）

$$SS_y = SS_{\hat{y}} + SS_e$$

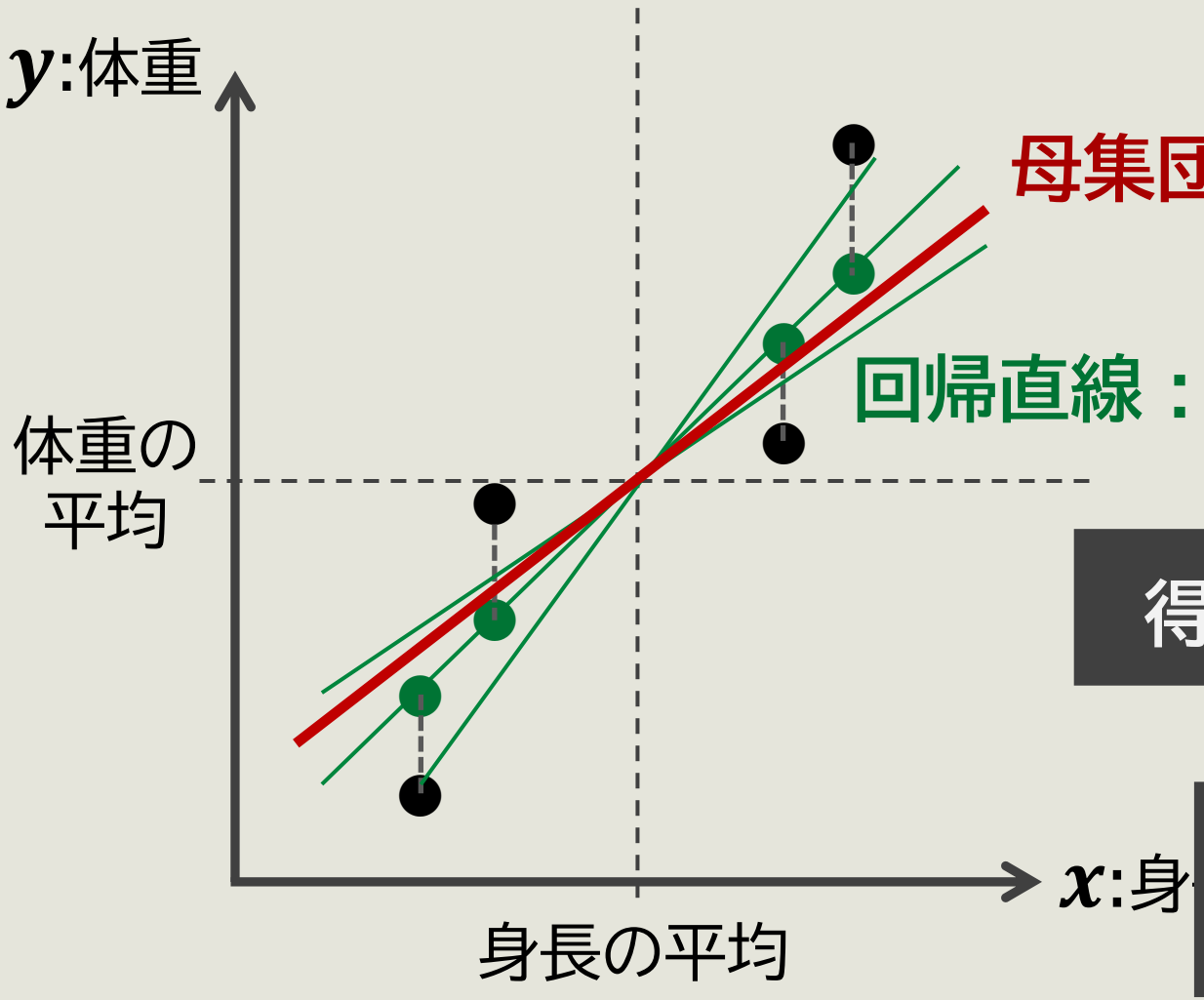
分散説明率：$SS_{\hat{y}}/SS_y = \cos^2\theta = r^2$

分散説明率（決定係数）：r^2

yの分散：$s_y{}^2 \rightarrow SS_y$	
$\times r^2$	$\times(1-r^2)$
$SS_{\hat{y}}$	SS_e

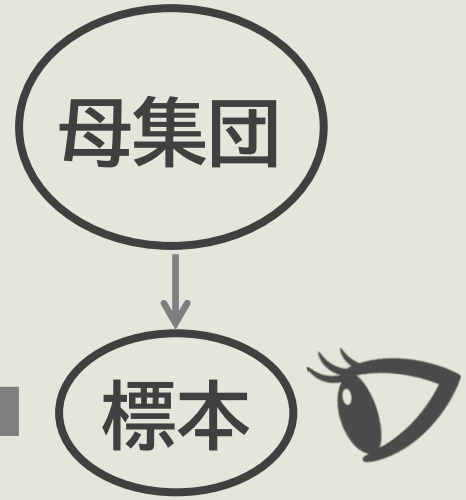
セクション4：単回帰分析の検定

知りたいのは母集団の真の回帰直線(母回帰直線)



母集団の真の回帰直線

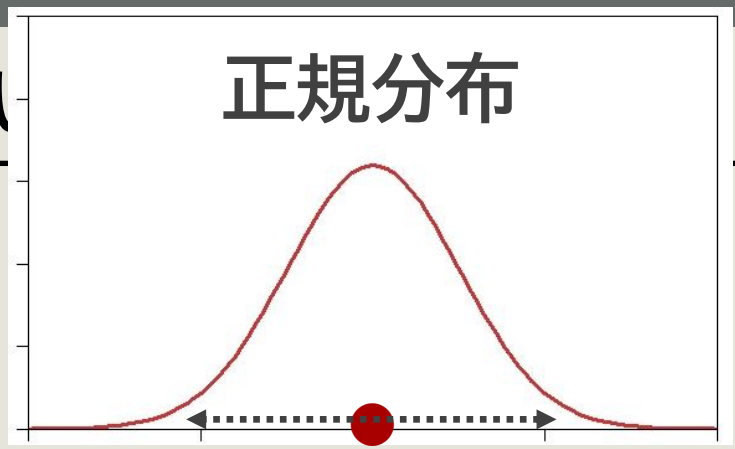
回帰直線：$\hat{y}_i = a + bx_i$

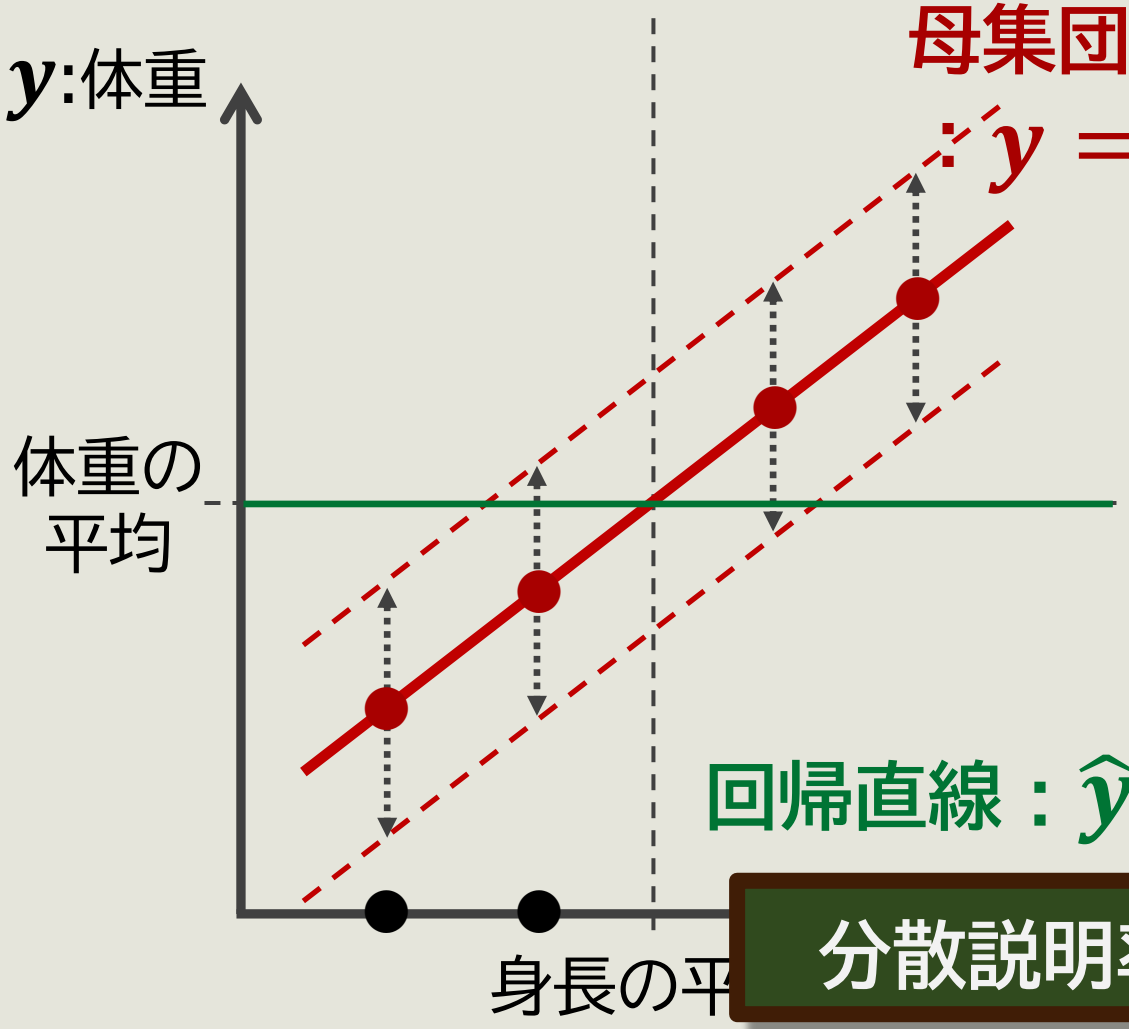


得られた回帰直線は信頼できるものか

(説明変数による)分散説明率を
確率的に検証(仮説検定)

説明変数が目的変数を何ら説明してくれない





母集団の真の回帰直線

$$y = \alpha + \beta x + \epsilon$$

重要な仮定：$\epsilon \sim N(0, {\sigma_\epsilon}^2)$

帰無仮説：$\beta = 0$

※xがyを何ら説明していない

回帰直線：$\hat{y}_i = a + bx_i = \overline{y} + b(x_i - \overline{x})$

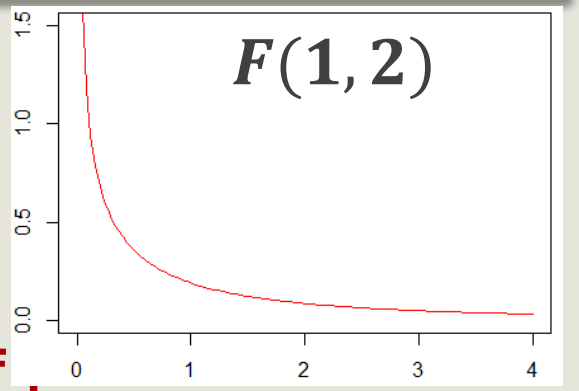
分散説明率：$SS_{\hat{y}}/SS_y = cos^2\,\theta = r^2 = 0$

仮説検定のために自由度の理解は必須

F分布：正規母集団からの独立な2つの平均平方和(SS)の比がしたがう確率分布

$$\frac{SS_1/n_1}{SS_2/n_2} \sim F(n_1, n_2)$$

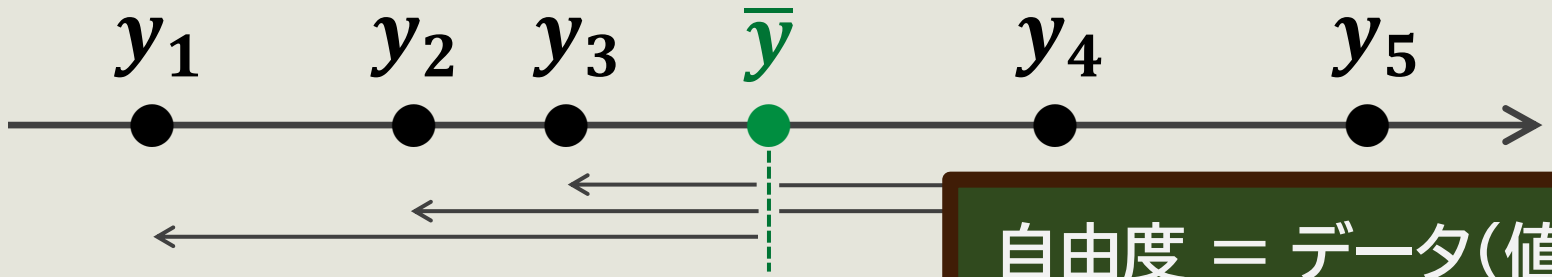
※n_1, n_2: 自由度



「平均」するために
「自由度」が必要

「データ(SS)」の「自由度」

《自由度のイメージ》



自由度は「5」だろうか？

自由度 ＝ データ(値)の数 － 推定された値の数

$$\bar{y} = \frac{y_1 + \cdots + y_5}{5} \Rightarrow y_5 = 5\bar{y} - (y_1 + \cdots + y_4)$$

仮説検定のために自由度の理解は必須

F分布：正規母集団からの独立な2つの平均平方和（SS）の比がしたがう確率分布

$$\frac{SS_1/n_1}{SS_2/n_2} \sim F(n_1, n_2)$$

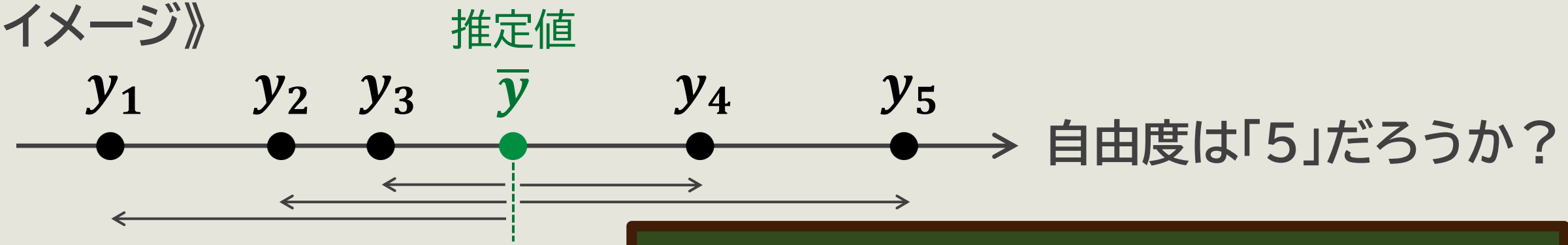
※n_1, n_2：自由度

➡ 「平均」するために「自由度」が必要

⬇

「データ（SS）」の「自由度」

《イメージ》



$$\bar{y} = \frac{y_1 + \cdots + y_5}{5} \Rightarrow y_5 = 5\bar{y} - (y_1 + \cdots + y_4)$$

自由度 ＝ データ（値）の数 － 推定された値の数

全体(実測値yの偏差)の平方和の自由度は？

全体平方和：$SS_y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$

自由度 ＝ データ(値)の数 － 推定された値の数

→1つ：$\bar{y}$

自由度 $= n - 1$

※$n = 3$のとき　$SS_y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2$

自由度 $= 3 - 1 = 2$

全体平方和の平均：$SS_y/(n-1)$

推定される母数の数だけ「自由」な値の数は減っていく

残差平方和：$SS_e = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$

※ $\hat{y}_i = a + bx_i$ $= (y_1 - a - bx_1)^2 + \cdots + (y_n - a - bx_n)^2$

自由度 ＝ データ(値)の数 － 推定された値の数

自由度 $= n - 2$

→ 2つ：a, b

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

$$b = r \times \frac{s_y}{s_x}$$

残差平方和の平均：$SS_e/(n-2)$

予測値平方和の自由度は全体平方和と残差平方和の自由度から求まる

自由度 ＝ データ(値)の数 － 推定された値の数

SS_yの自由度 $= n - 1$

SS_eの自由度 $= n - 2$

→1つ：$\bar{y}$

→2つ：a, b

$$SS_y = SS_{\hat{y}} + SS_e$$

SS_yの自由度 ＝ $SS_{\hat{y}}$の自由度 ＋ SS_eの自由度

$SS_{\hat{y}}$の自由度 $= (n-1) - (n-2) = 1$

予測値平方和の平均：$SS_{\hat{y}}/1$

平均平方(の比)がわかればF分布による仮説検定ができる

F分布:正規母集団からの独立な2つの平均平方和(SS)の比がしたがう確率分布

重要な仮定:$\epsilon \sim N(0, {\sigma_\epsilon}^2)$ → eや$\hat{y}$が正規分布

帰無仮説:$\beta = 0$
※xがyを何ら説明していない

→ $r = 0 \rightarrow SS_{\hat{y}}$と$SS_e$は互いに独立
※$SS_y = SS_{\hat{y}} + SS_e$

$$\frac{SS_{\hat{y}}/1}{SS_e/(n-2)} \sim F(1, n-2)$$

平均平方和(の比)がわかればF分布による仮説検

$$\frac{SS_1/n_1}{SS_2/n_2} \sim F(n_1, n_2)$$
※n_1, n_2: 自由度

F分布:正規母集団からの独立な2つの平均平方和(SS)の比がしたがう確率分布

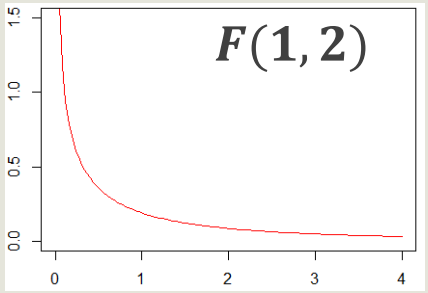
重要な仮定：$\epsilon \sim N(0, {\sigma_\epsilon}^2)$ ⇒ $y \sim N(\alpha + \beta x, {\sigma_\epsilon}^2)$ 正規分布

独立？？

$$SS_{\hat{y}} = SS_y \times r^2$$
$$SS_e = SS_y \times (1 - r^2)$$
⇒ $$SS_{\hat{y}} = \frac{r^2}{1 - r^2} \times SS_e$$

帰無仮説：$\beta = 0$ ⇒ 分散説明率：$SS_{\hat{y}}/SS_y = r^2 = 0$

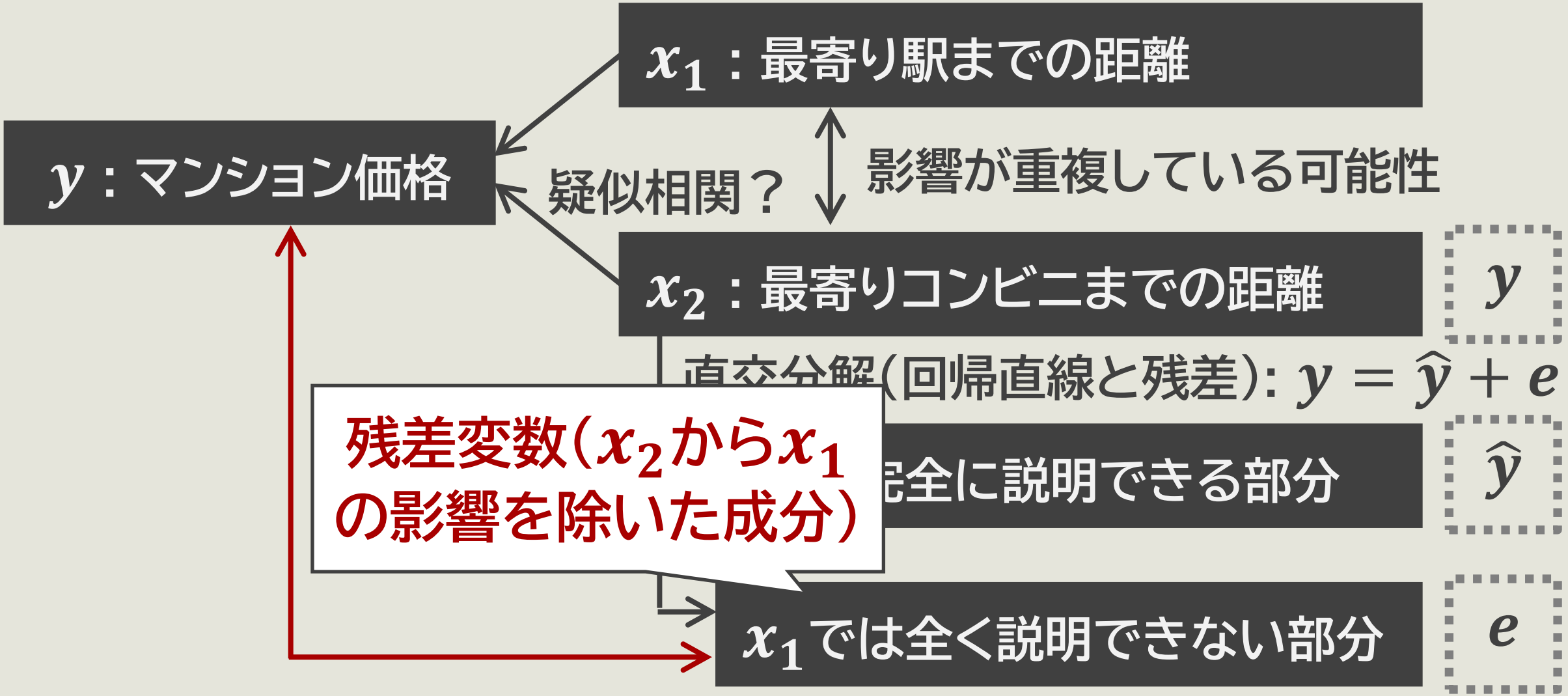
$$\frac{SS_{\hat{y}}/1}{SS_e/(n-2)} \sim F(1, n-2)$$



$SS_{\hat{y}}$とSS_eは互いに独立

セクション5：重回帰分析

他の変数の影響による疑似相関を回避したい



他の変数の影響を除いて残った「部分」との相関を考える

y：マンション価格

x_1：最寄り駅までの距離

x_2：最寄りコンビニまでの距離

疑似相関？

影響が重複している可能性

直交分解（回帰直線と残差）：$y = \hat{y} + e$

y

$\hat{y}$

明できる部分

e

x_1では全く説明できない部分

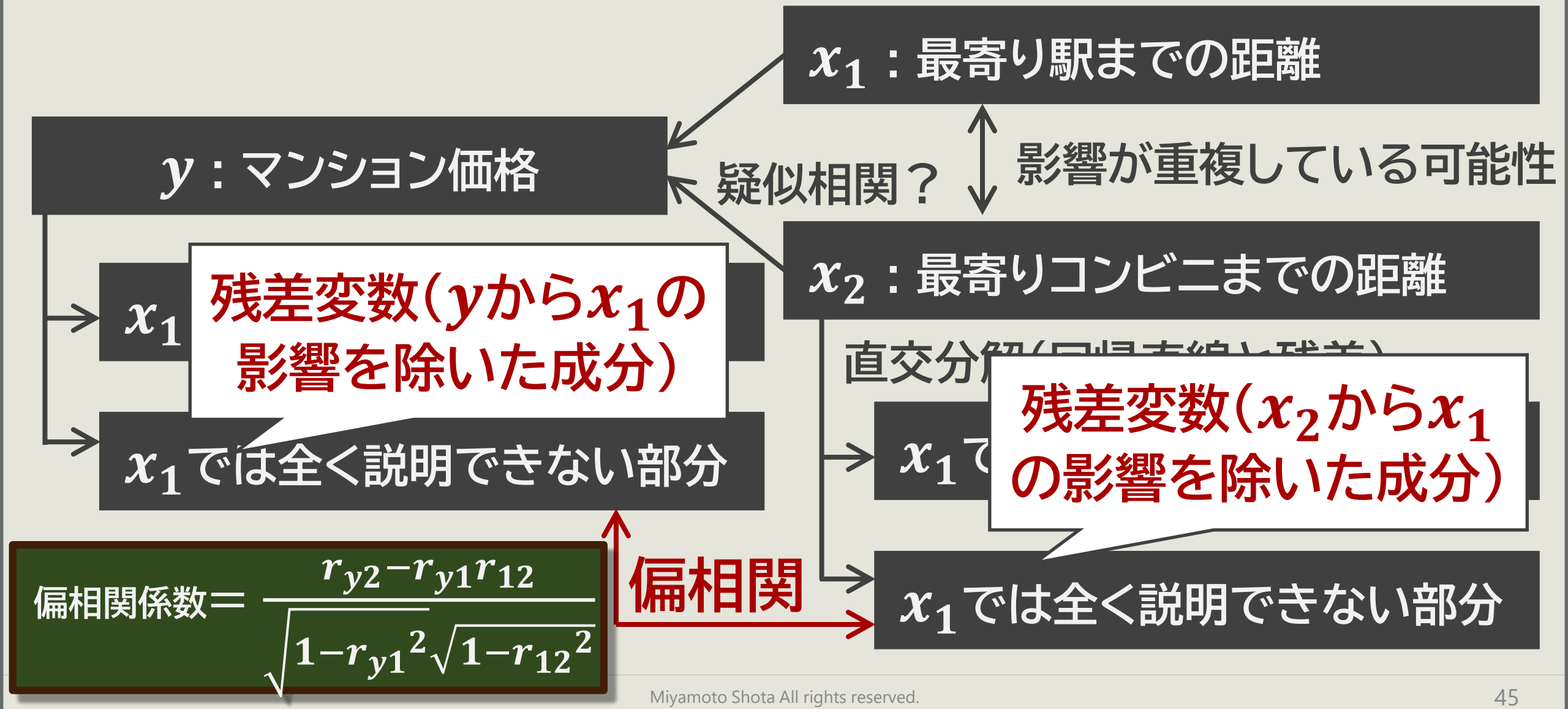
残差変数（x_2からx_1の影響を除いた成分）

部分相関係数＝ $\dfrac{r_{y2}-r_{y1}r_{12}}{\sqrt{1-{r_{12}}^2}}$

r_{y1}：yとx_1の相関係数
r_{y2}：yとx_2の相関係数
r_{12}：x_1とx_2の相関係数

部分相関

他の変数の影響を除いて残った「部分同士」での相関を考える



$$\text{偏相関係数} = \frac{r_{y2} - r_{y1} r_{12}}{\sqrt{1 - {r_{y1}}^2}\sqrt{1 - {r_{12}}^2}}$$

他の説明変数の影響を除いたその説明変数「独自の成分」にかかる係数

x_1：最寄り駅までの距離

y：マンション価格

疑似相関？ ↕ 影響が重複している可能性

x_2：最寄りコンビニまでの距離

y

直交分解（回帰直線と残差）：$y = \hat{y} + e$

$$b_2 = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - {r_{12}}^2}} \times \frac{s_y}{s_{(x_2|x_1)}} = \frac{r_{y2} - r_{y1}r_{12}}{1 - {r_{12}}^2} \times \frac{s_y}{s_{x_2}}$$

x_1で完全に説明できる部分

x_1の影響を受けないx_2の「独自の成分」：$x_2|x_1$

x_1では全く説明できない部分

e

$$\hat{y} = a + b_2 \times x_2|x_1$$

↓

b_2：偏回帰係数

偏回帰係数について単位の影響を受けないように変換

y(マンション価格)とx_1(最寄駅距離)、x_2(最寄コンビニ距離)について…

$$y\text{と}x_2\text{の部分相関係数} = \frac{r_{y2}-r_{y1}r_{12}}{\sqrt{1-{r_{12}}^2}}$$

$$y\text{と}x_2\text{の偏相関係数} = \frac{r_{y2}-r_{y1}r_{12}}{\sqrt{1-{r_{y1}}^2}\sqrt{1-{r_{12}}^2}}$$

r_{y1}：yとx_1の相関係数
r_{y2}：yとx_2の相関係数
r_{12}：x_1とx_2の相関係数

$$x_2|x_1\text{にかかる偏回帰係数} = \frac{r_{y2}-r_{y1}r_{12}}{1-{r_{12}}^2} \times \frac{s_y}{s_{x_2}}$$

分子は
すべて同じ

$$x_2|x_1\text{にかかる標準偏回帰係数} = \frac{r_{y2}-r_{y1}r_{12}}{1-{r_{12}}^2}$$

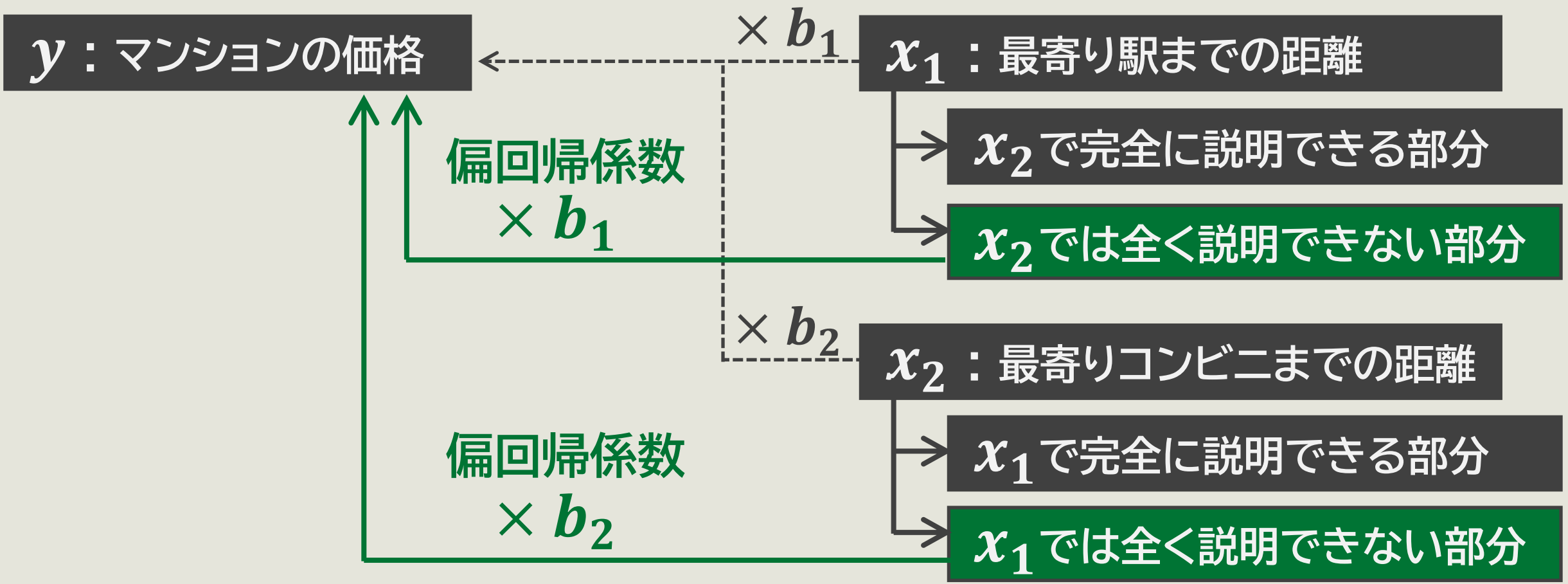
その説明変数「独自の部分」（残差変数）にかかる係数

- y：マンション価格
  - x_1：最寄り駅までの距離
  - x_2：最寄りコンビニまでの距離
    - x_1で完全に説明できる部分
    - x_1では全く説明できない部分

- y：売上
  - x_1：店舗面積
  - x_2：取扱い商品点数
    - x_1で完全に説明できる部分
    - x_1では全く説明できない部分

各々の説明変数の「独自の部分」と目的変数との関係を調べる

重回帰式 :$y = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$



（偏）微分して切片 a と複数の回帰係数 b についての連立方程式を解く

重回帰式：$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$

y：マンション価格（万円）
x_1：最寄駅距離（km）
x_2：部屋の広さ（㎡）

i	y	x_1	x_2
1	4,400	1.2	65
2	4,800	1.0	65
3	5,600	0.4	60
4	6,600	0.6	70

適合の悪さ：$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

$= \sum_{i=1}^{N}(y_i - a - b_1x_{1i} - b_2x_{2i})^2$

$= (4400 - a - 1.2b_1 - 65b_2)^2 + \cdots$

$\quad +(6600 - a - 0.6b_1 - 70b_2)^2$

$= \bullet a^2 + \blacksquare {b_1}^2 + \blacktriangle {b_2}^2 + \cdots$

（偏）微分して切片 a と複数の回帰係数 b についての連立方程式を解く

重回帰式：$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$

y：マンション価格（万円）
x_1：最寄駅距離（km）
x_2：部屋の広

適合の悪さ $= \sum_{i=1}^{N}(y_i - \hat{y})^2$ → 最小

$$b_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - {r_{12}}^2} \times \frac{s_y}{s_{x1}} \quad b_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - {r_{12}}^2} \times \frac{s_y}{s_{x2}}$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

r_{y1}：yとx_1の相関係数
r_{y2}：yとx_2の相関係数
r_{12}：x_1とx_2の相関係数

i	y		
1	4,400		
2	4,800		
3	5,600	0.4	60
4	6,600	0.6	70

$$\hat{y} = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2)$$

実測値と予測値の(重)相関係数

重回帰式：$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$

単回帰分析においては…

分散説明率(決定係数)：r^2

yの分散：${s_y}^2$

$\times r^2$	$\times (1 - r^2)$
xで完全に説明できる部分	xでは全く説明できない部分

r：yと$x(\hat{y})$の相関係数

重回帰分析においては…

分散説明率(決定係数)：R^2

yの分散：${s_y}^2$

$\times R^2$	$\times (1 - R^2)$
$x_{i\ldots}$で完全に説明できる部分	$x_{i\ldots}$では全く説明できない部分

R：yと$\hat{y}$の相関係数(重相関係数)

セクション6：重回帰分析の視覚的理解

ベクトル表現のおさらい

変数のベクトル表現

「偏差」変数をベクトルで表現すると都合が良い！

$\vec{x}$: xベクトル(向きと大きさを持つ)

$\vec{x} = (168, 176)$
1人目 2人目

$\vec{x}$ の大きさ: $\|\vec{x}\| = \sqrt{168^2 + 176^2}$

$\vec{x} = (168 - \bar{x}, 176 - \bar{x})$
$= (168 - 172, 176 - 172)$
$= (-4, 4)$ 平均からの偏差ベクトル

$\vec{x}$ の大きさ: $\|\vec{x}\| = \sqrt{(-4)^2 + 4^2}$

xの分散 : $s_x{}^2 = \|\vec{x}\|^2/n$
xの標準偏差 : $s_x = \|\vec{x}\|/\sqrt{n}$

ベクトルの内積と共分散

ベクトルの内積を使うとさらに都合が良い！

1人目 2人目
身長 : $\vec{x} = (168, 176)$ → 身長の偏差 : $\vec{x} = (168 - \bar{x}, 176 - \bar{x}$
体重 : $\vec{y} = (\ 70,\ 74\ )$ → 体重の偏差 : $\vec{y} = (70 - \bar{y}, 74 - \bar{y})$

《偏差ベクトルの内積》 1人目 2人目
$\vec{x} \cdot \vec{y} = (168 - \bar{x})(70 - \bar{y}) + (176 - \bar{x})(74 - \bar{y})$

共分散 : $s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$

偏差ベクトルの内積 : $\vec{x} \cdot \vec{y} = n \times s_{xy}$

相関係数 $\cos\theta$ ①

相関係数 $= \cos\theta$

(偏差)ベクトルの内積 : $\vec{x} \cdot \vec{y} = n \times s_{xy}$

《(偏差)ベクトルの内積の(もうひとつの)定義》
$\vec{x} \cdot \vec{y} = \|\vec{x}\| \times \|\vec{y}\| \times cos\,\theta$

xの標準偏差 : $s_x = \|\vec{x}\|/\sqrt{n} \rightarrow \|\vec{x}\| = \sqrt{n}s_x$

$= \sqrt{n}s_x \times \sqrt{n}s_y \times cos\,\theta = ns_xs_y\,cos\,\theta = ns_{xy}$

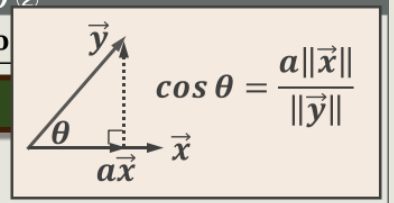
$cos\,\theta = s_{xy}/s_xs_y = r$: 相関係数

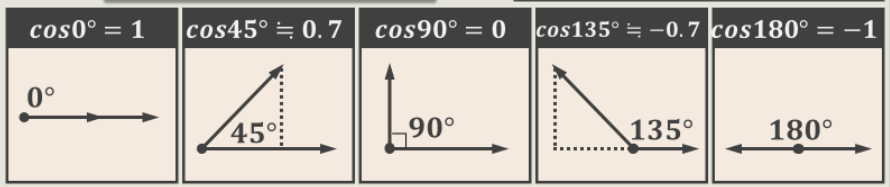
相関係数 $\cos\theta$ ②

相関係数 $= co$

$cos\,\theta = s_{xy}/s_xs_y = r$: 相関係数

$-1 \le cos\,\theta = r \le 1$





2つのベクトルが直角のときに無相関となる。
2つのベクトルの向きが同じ(または正反対)のときに完全相関となる。

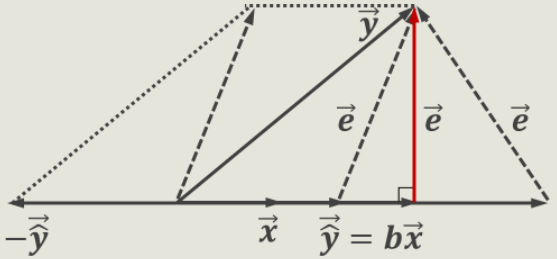
最小二乗法のベクトル的意味

最小二乗法は残差ベクトル $\vec{e}$ を最短にする

残差ベクトル : $\vec{e} = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots) = \vec{y} - \hat{\vec{y}}$

適合の悪さ : $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \|\vec{e}\|^2$

最小二乗法は残差ベクトル$\vec{e}$の大きさを最小にする…



残差ベクトル$\vec{e}$の大きさが最小となるとき、
$\hat{\vec{y}}$と$\vec{e}$は直角(直交する)

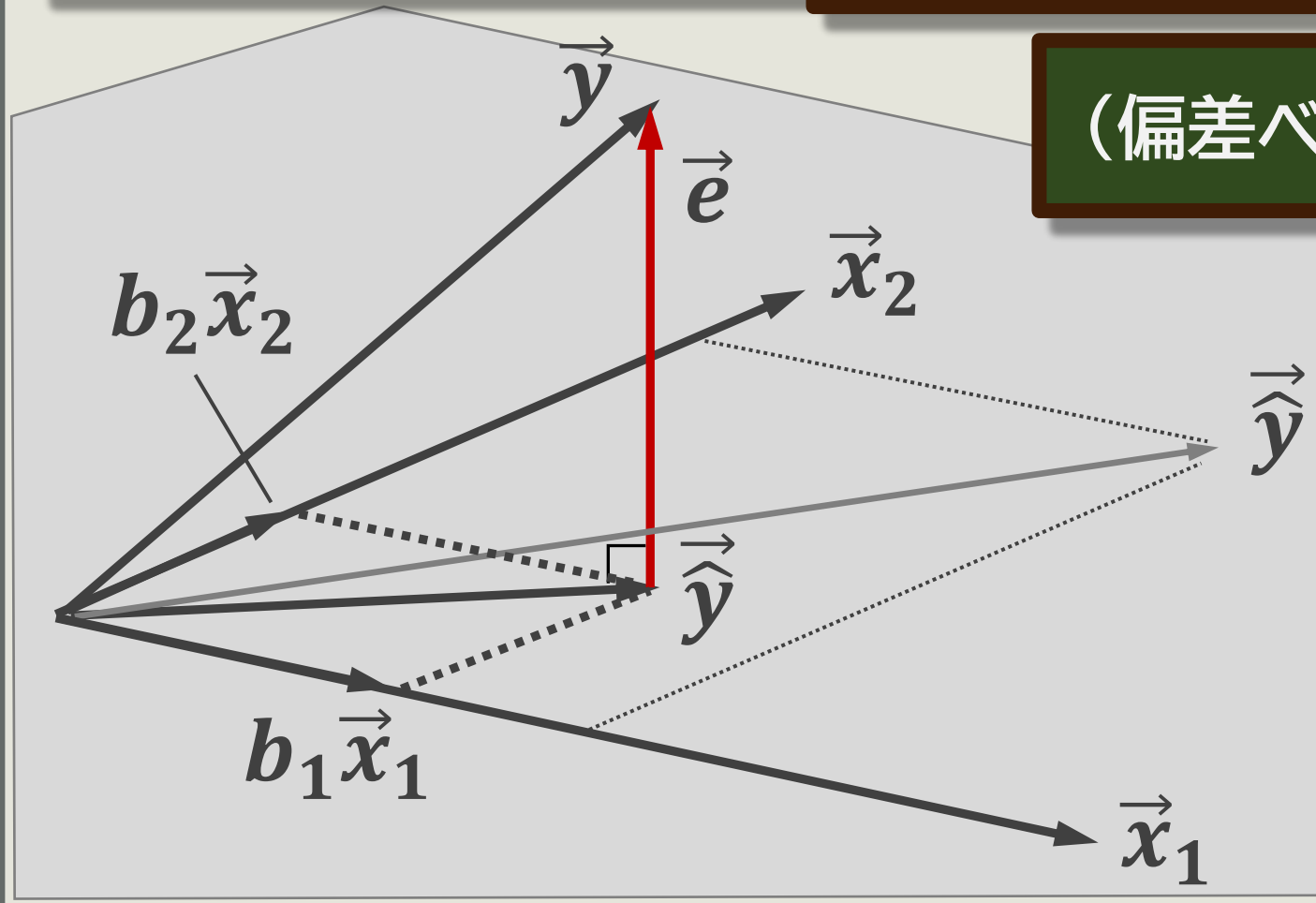
重回帰分析のベクトル表現

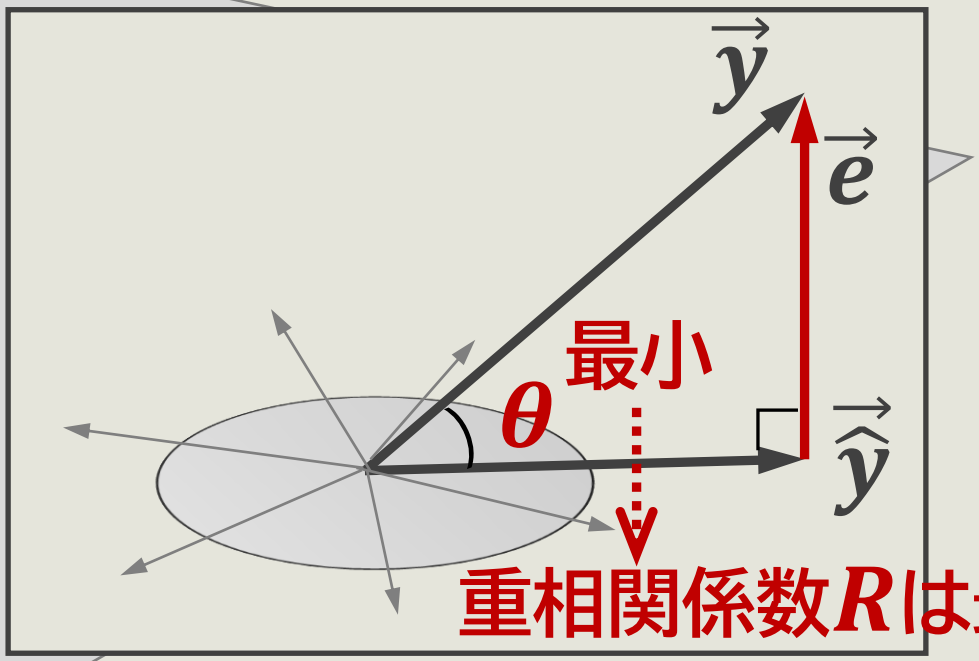
最小二乗法の予測値ベクトル $\vec{\hat{y}}$ は、実測値ベクトル $\vec{y}$ との距離が最短のところ

重回帰式：$\hat{y} = a + b_1$

$$\hat{y} = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2)$$

（偏差ベクトルで）$\vec{\hat{y}} = b_1\vec{x}_1 + b_2\vec{x}_2$

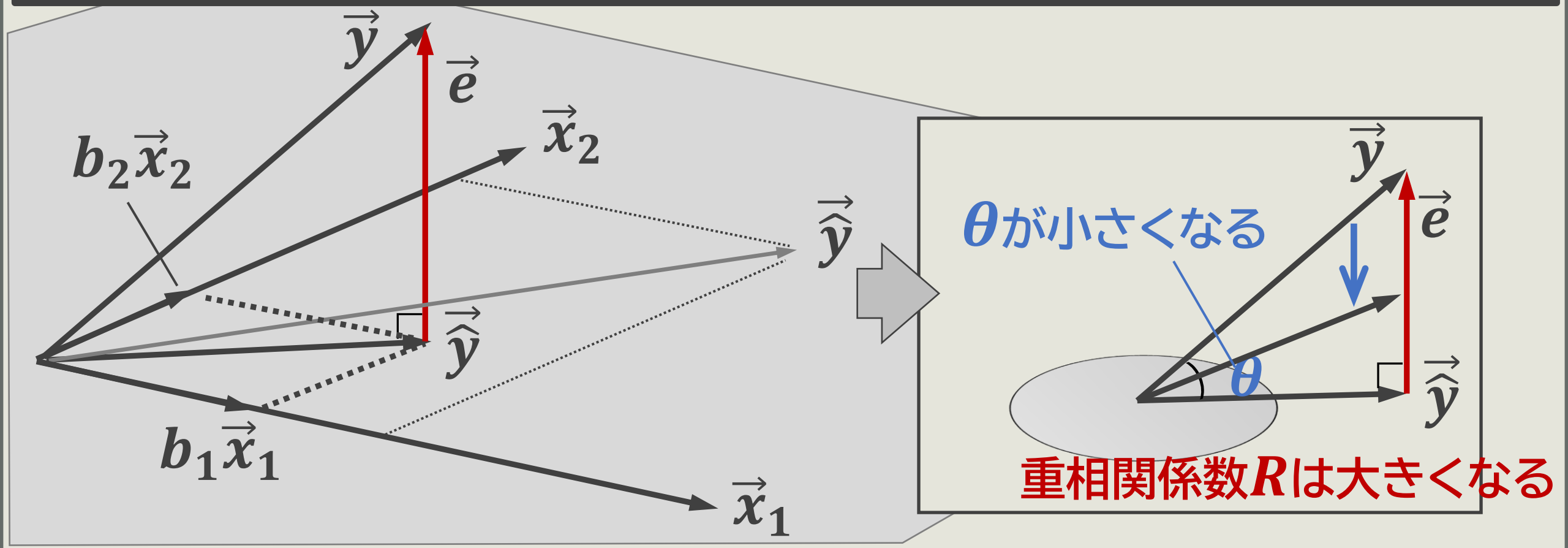




重相関係数の変動

予測値ベクトル$\vec{\hat{y}}$と実測値ベクトル$\vec{y}$との(重)相関を大きくするには？

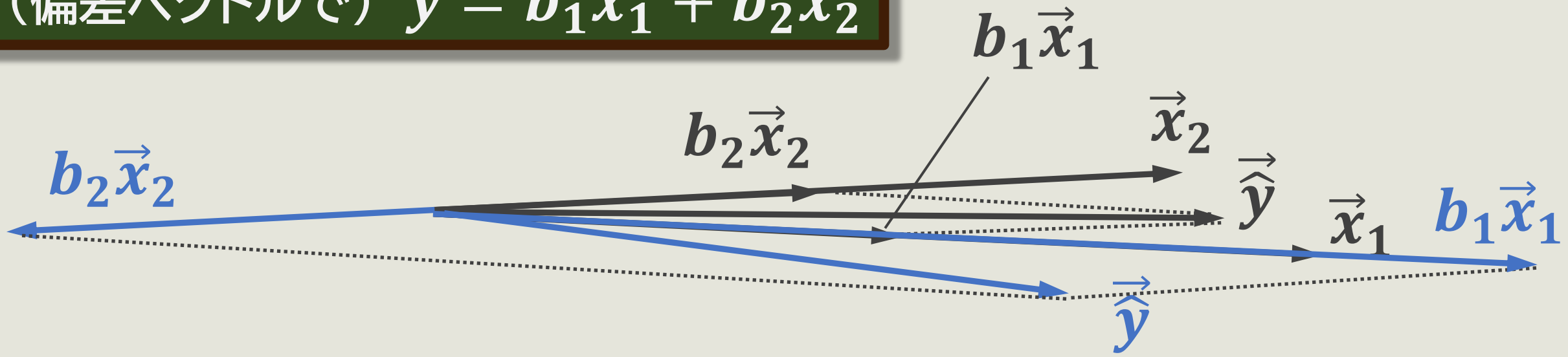
- ✓ $\vec{y}$と$\vec{x}_1, \vec{x}_2$の成す角が小さくなる($\vec{y}$と$\vec{x}_1, \vec{x}_2$の相関が大きくなる)とき
- ✓ $\vec{x}_1$と$\vec{x}_2$の成す角が大きくなる($\vec{x}_1$と$\vec{x}_2$の相関が小さくなる)とき



説明変数同士の相関が大きいと回帰係数の推定が不安定になる

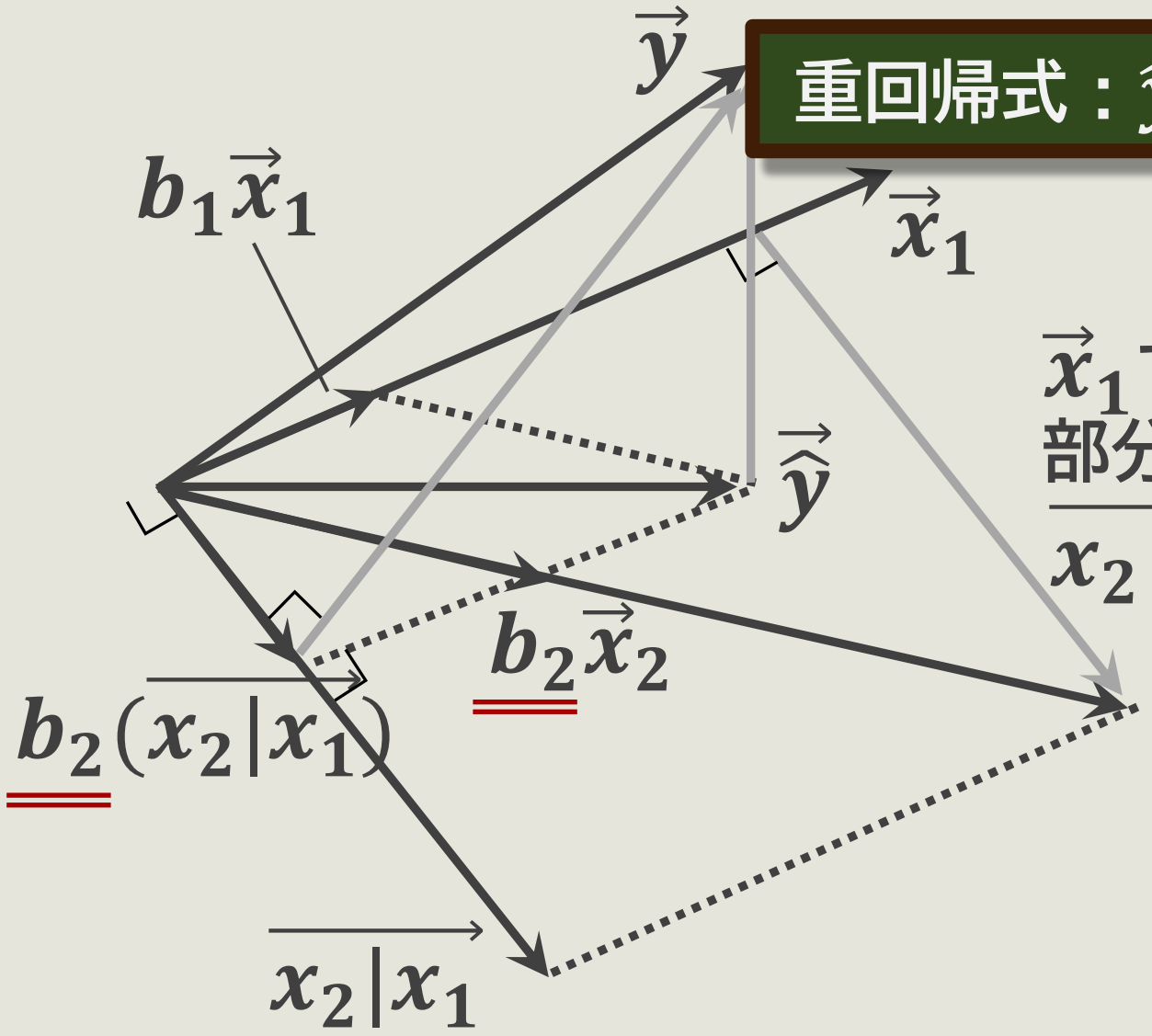
$$\hat{y} = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2)$$

（偏差ベクトルで）$\vec{\hat{y}} = b_1\vec{x}_1 + b_2\vec{x}_2$



✓ $\vec{x}_1$と$\vec{x}_2$の成す角が小さくなる（$\vec{x}_1$と$\vec{x}_2$の相関が大きくなる）とき
回帰係数b_1, b_2のブレが大きくなる

「他の説明変数を固定したときの」という表現の意味



重回帰式：$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$

《偏回帰係数の説明》

説明変数が1単位変化すると目的変数がどれだけ変化するかを示す指標

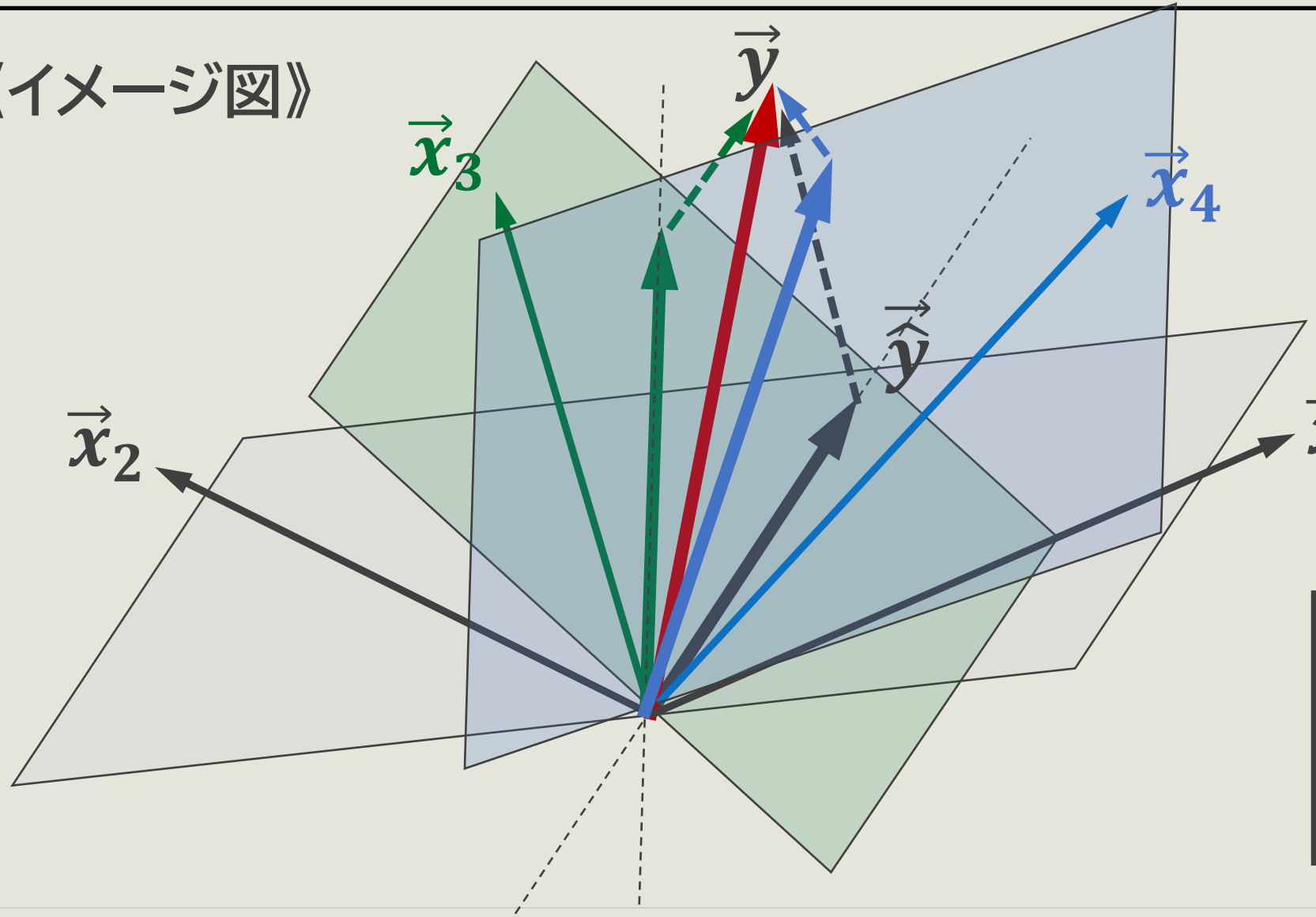
（重回帰分析の枠組みで）
「他の説明変数を固定したときに…」

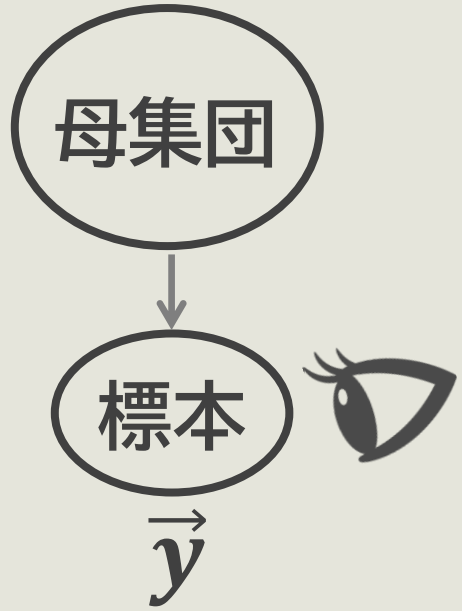
（単回帰分析の枠組みで）
「他の説明変数の影響を除外した残差変数において…」

セクション7：重回帰分析の検定

説明変数の数が多いほど$\vec{y}$に寄り過ぎる（重相関係数が大きくなる）

《イメージ図》





$\vec{y}$に寄り過ぎる

↓

“標本”に寄り過ぎる

決定係数は標本に寄り過ぎる度合いを調整できていない

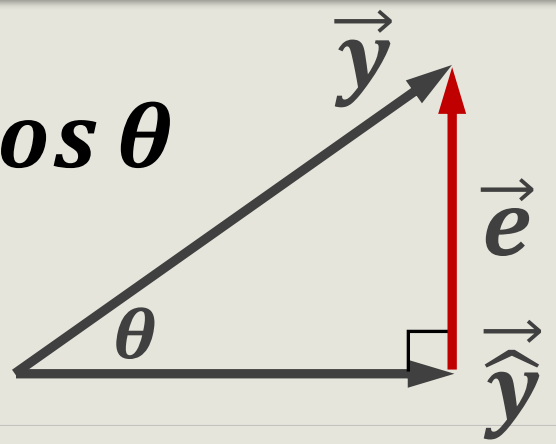
分散説明率(決定係数) : R^2

yの分散 : ${s_y}^2$

$\times R^2$ | $\times (1 - R^2)$

R : yと$\hat{y}$の相関係数(重相関係数)

$R = \cos\theta$



分散説明率(決定係数)

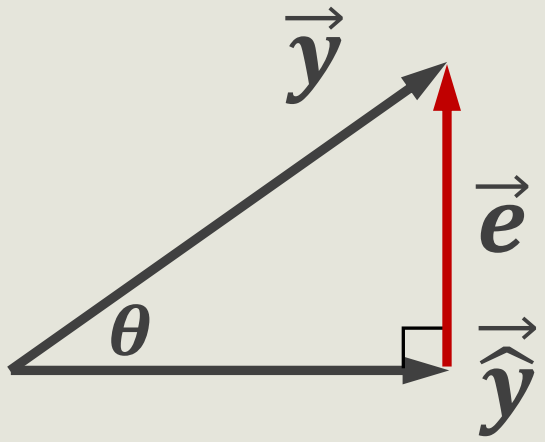
: $R^2 = \cos^2\theta$

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

SS : 平方和(*Sum of Squares*)

$$R^2 = \frac{SS_{\hat{y}}}{SS_y} = 1 - \frac{SS_e}{SS_y}$$

標本に寄り過ぎる分を自由度で調整してあげる



分散説明率(決定係数)

$$: R^2 = \frac{SS_{\hat{y}}}{SS_y} = 1 - \frac{SS_e}{SS_y} = 1 - \frac{SS_e/n}{SS_y/n}$$

自由度 ＝ データ(値)の数 － 推定された値の数

自由度調整済み決定係数

$$: {R_{adj}}^2 = 1 - \frac{SS_e/(n-p-1)}{SS_y/(n-1)}$$

$$= 1 - \frac{SS_e}{SS_y} \times \frac{n-1}{n-p-1}$$

nよりpが大きい分だけ
残った分散が大きくなるように
(R^2が小さくなるように)補正

単回帰分析の検定を応用してモデル全体を検定

F分布：正規母集団からの独立な2つの平均平方和(SS)の比がしたがう確率分布

重回帰モデル：$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$

重要な仮定：$\epsilon \sim N(0, {\sigma_\epsilon}^2)$ ⇒ $y \sim N(\alpha + \beta_1 x_1 + \cdots, {\sigma_\epsilon}^2)$

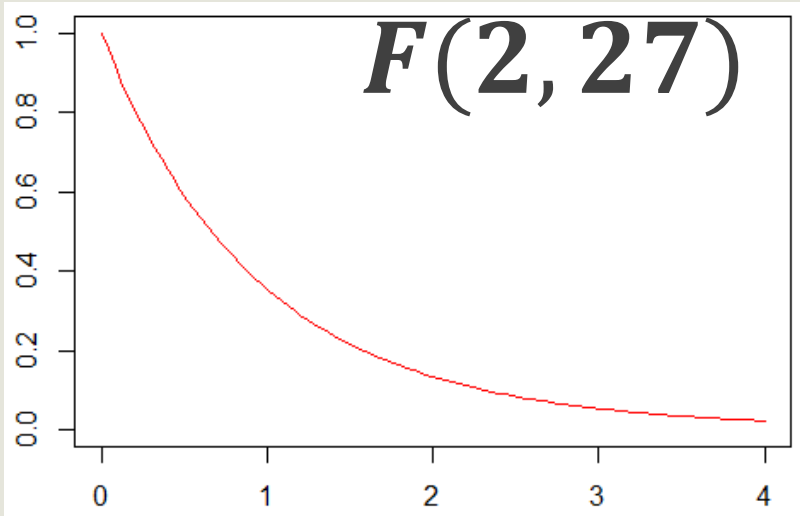
帰無仮説：$\beta_1 = \beta_2 = \cdots = \beta_p = 0$

※yを何ら説明していない

$SS_{\hat{y}}$とSS_eは互いに独立

$$\frac{SS_{\hat{y}}/p}{SS_e/(n-p-1)} \sim F(p, n-p-1)$$

（※p: 説明変数の数）



偏回帰係数の分布の性質からt分布により検定

重回帰モデル：$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$

重要な仮定：$\epsilon \sim N(0, {\sigma_\epsilon}^2)$　※yも正規分布

偏回帰係数の検定のために… $b_1 \sim$ ■(●, ▲) を知りたい！

正規分布：N

平均＝真の偏回帰係数：β_1と一致

分散${\sigma_{b1}}^2 = \dfrac{{\sigma_\epsilon}^2}{n{s_{x1}}^2(1-R^2)}$

${\sigma_\epsilon}^2$ ？

$\dfrac{b_1 - \beta_1}{\sigma_{b1}}$ が標準正規分布にしたがう…

t分布

代用

${s_e}^2 = \dfrac{SS_e}{N-p-1}$

以上