第2章
最小2乗法と決定係数

『入門　実践する計量経済学』
PPT

- 回帰モデル
- データの整理
- 最小2乗法
- 回帰直線と予測
- 残差の性質
- 決定係数

回帰モデル

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

--- Yは被説明変数

--- $X_1, X_2 \ldots, X_K$は、K個の説明変数

--- 関心はパラメータ($\alpha, \beta_1, \beta_2, \ldots, \beta_K$)の推定にある

---α : 数学的切片、もしくは$X = 0$のときのYの値

---$\beta_1, \beta_2, \ldots, \beta_K$: 回帰係数

--- 回帰係数β_1

X_1以外の説明変数を一定として、X_1を1単位増やしたとき、Yが何単位変化するか

[証明] $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$ から、

X_1を1単位増やすと

$$Y' = \alpha + \beta_1 (X_1 + 1) + \beta_2 X_2 + \cdots + \beta_K X_K$$

このとき、Yの変化分はβ_1となる

$$Y' - Y = [\alpha + \beta_1 (X_1 + 1) + \beta_2 X_2 + \cdots + \beta_K X_K]$$
$$- [\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K] = \beta_1$$ [終]

--- 説明変数が1個なら単回帰モデル(K=1)、2個以上なら重回帰モデルと呼ばれる($2 \leq K$)

--- 本章では単回帰モデルを扱う

$$Y_i = \alpha + \beta X_i \quad (i = 1,2, \ldots, n)$$

データの整理

・データ

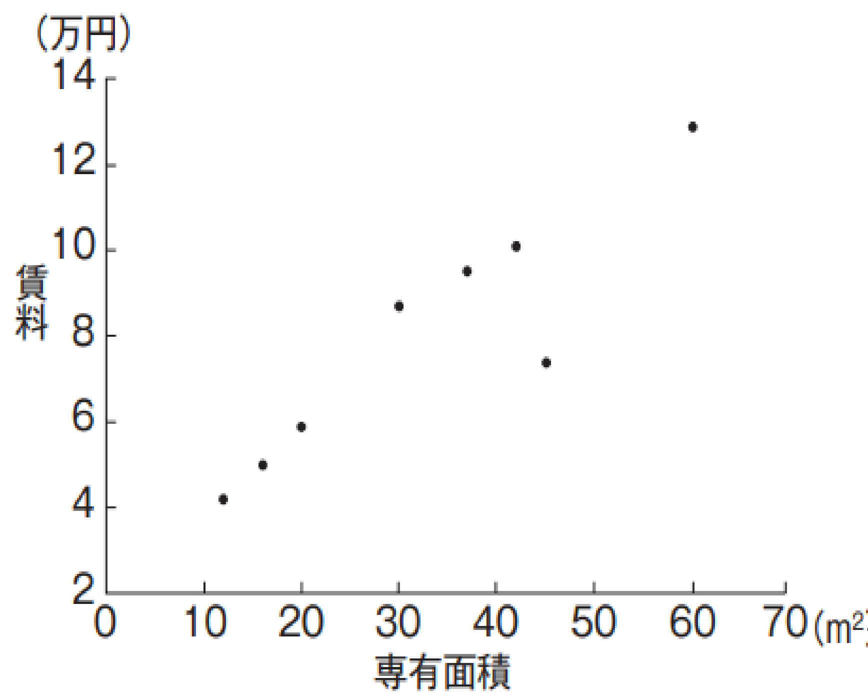
$$(Y_1, X_1)、(Y_2, X_2)、\ldots、(Y_n, X_n)$$

--- サンプルサイズはn個

- 表２-２　賃料と専有面積

i	賃料(万円) Y	面積(m²) X
1	4.2	12
2	5.0	16
3	5.9	20
4	7.4	45
5	8.7	30
6	9.5	37
7	10.1	42
8	12.9	60

- 図２-１　賃料と専有面積の関係を表す散布図



特性値の復習

- 標本平均

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

--- データの中心を測る代表的指標

- 標本分散

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

--- データのばらつきを測る指標

- 標本標準偏差

$$s_X = \sqrt{s_X^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

--- データのばらつきを測る指標

- 標本共分散

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$$

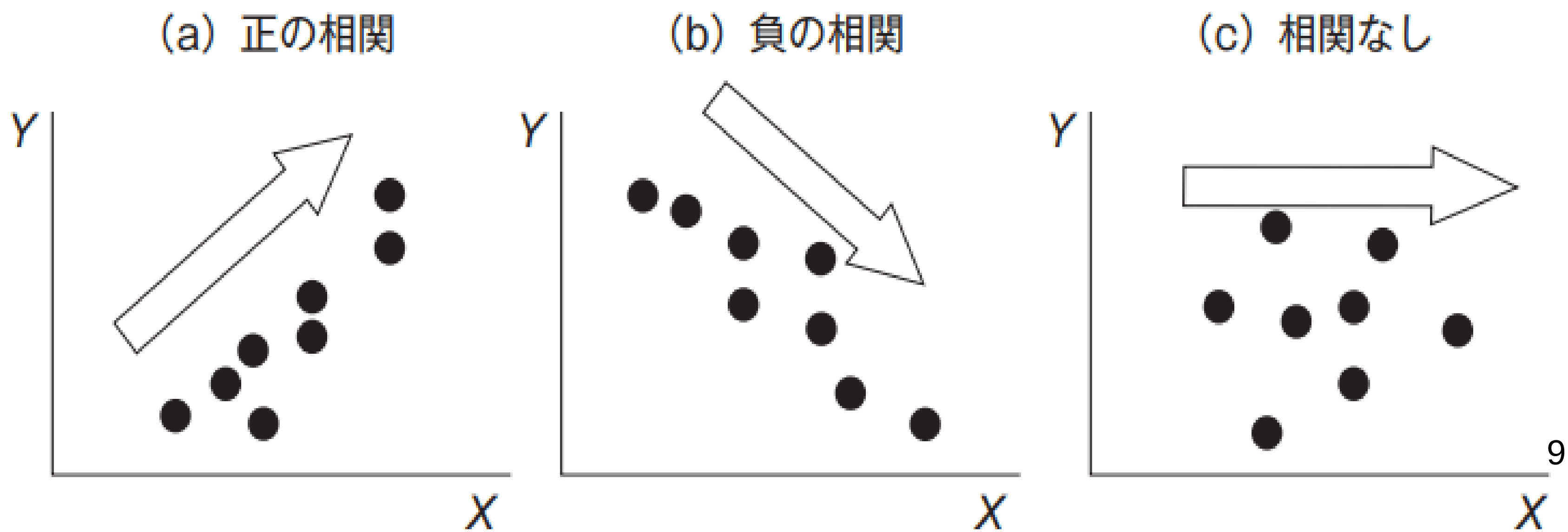
--- $s_{XY} > 0$なら正の相関、$s_{XY} < 0$なら負の相関

- 標本相関係数

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{s_{XY}}{s_X s_Y}$$

--- −1から+1の間で定義される

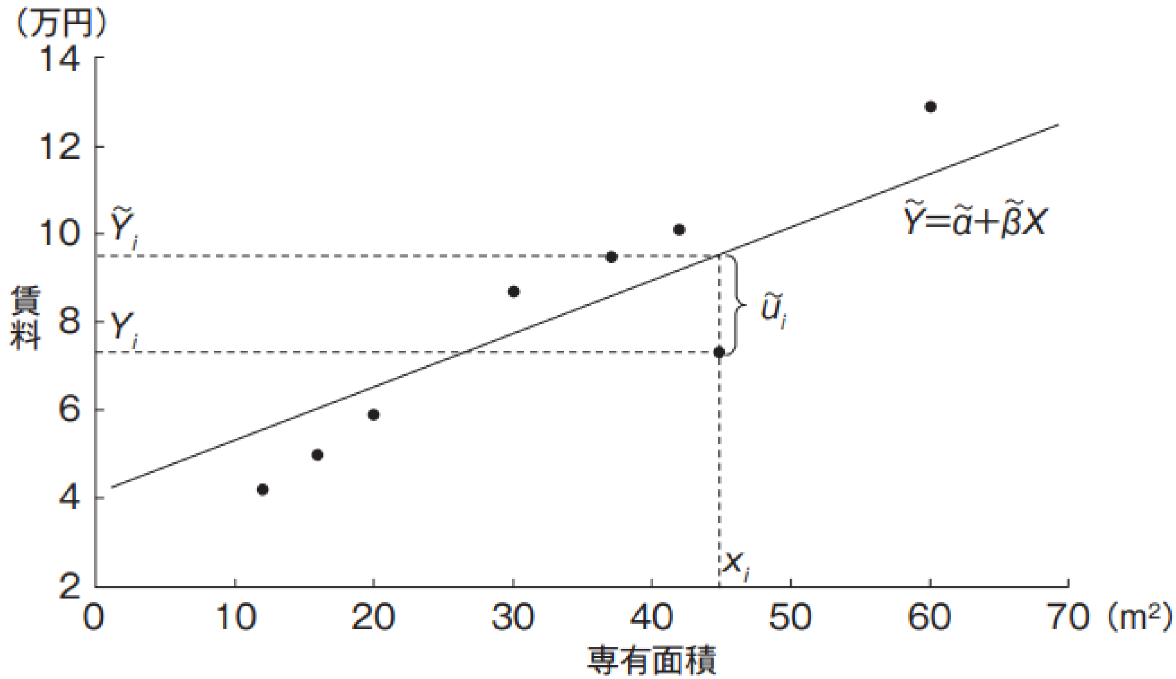
図 2 - 2　相関関係のイメージ



最小2乗法

当てはまりの良い線をどうやって引いたらよいか？

図 2 - 4　線形関係と残差



- データは$(X_1, Y_1)(X_2, Y_2) \ldots (X_n, Y_n)$
- 残差$\tilde{u}_i$が「小さい」ほど、当てはまりが良い

$$\tilde{u}_i = Y_i - \tilde{Y}_i = Y_i - (\tilde{\alpha} + \tilde{\beta} X_i)$$

残差を小さくするように線を引けばよい！

・残差の総和

$$\sum_{i=1}^{n} \tilde{u}_i = \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta} X_i)$$

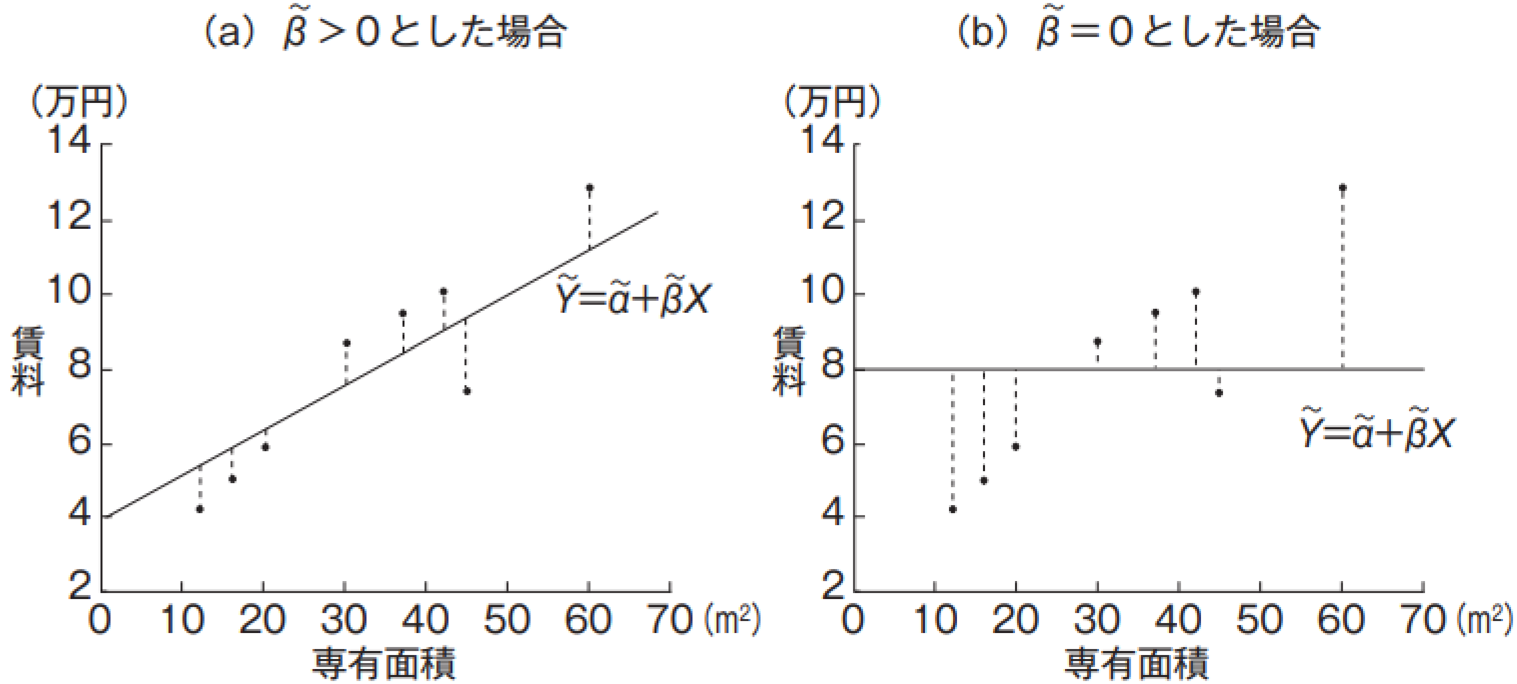
--- プラスとマイナスで打ち消し合ってしまう

・残差2乗和

$$\sum_{i=1}^{n} \tilde{u}_i^2 = \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta} X_i)^2$$

--- 全てプラスになるため打ち消し合わない

図２-５　全物件の残差



- 最小2乗推定量（OLS推定量）

$\tilde{\alpha}, \tilde{\beta}$を残差2乗和を最小にするように選ぶ（$\hat{\alpha}$と$\hat{\beta}$と表記）

$$\boxed{\begin{gathered}\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \\ \hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\end{gathered}}$$

$\hat{\beta}$の意味を考える

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

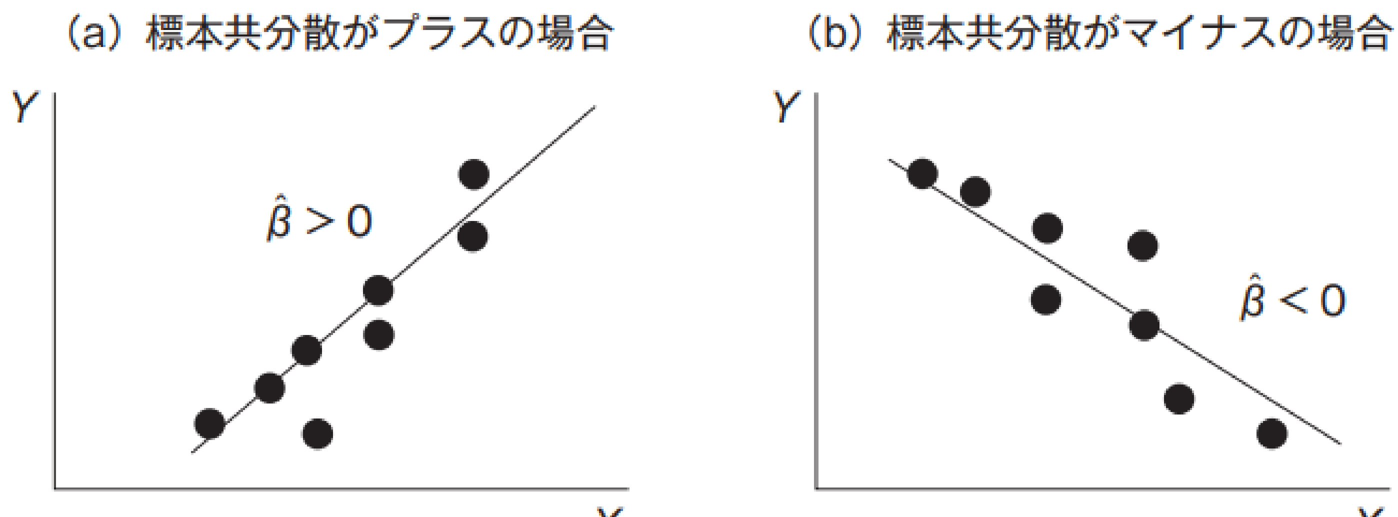
$$= \frac{\boxed{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})\,/(n-1)}}{\sum_{i=1}^{n}(X_i - \bar{X})^2\,/(n-1)}$$

--- 分子はXとYの標本共分散、分母はXの標本分散、

⇒傾きの符号条件は標本共分散にだけ依存する

--- 標本共分散が正（負）なら、傾きは正（負）

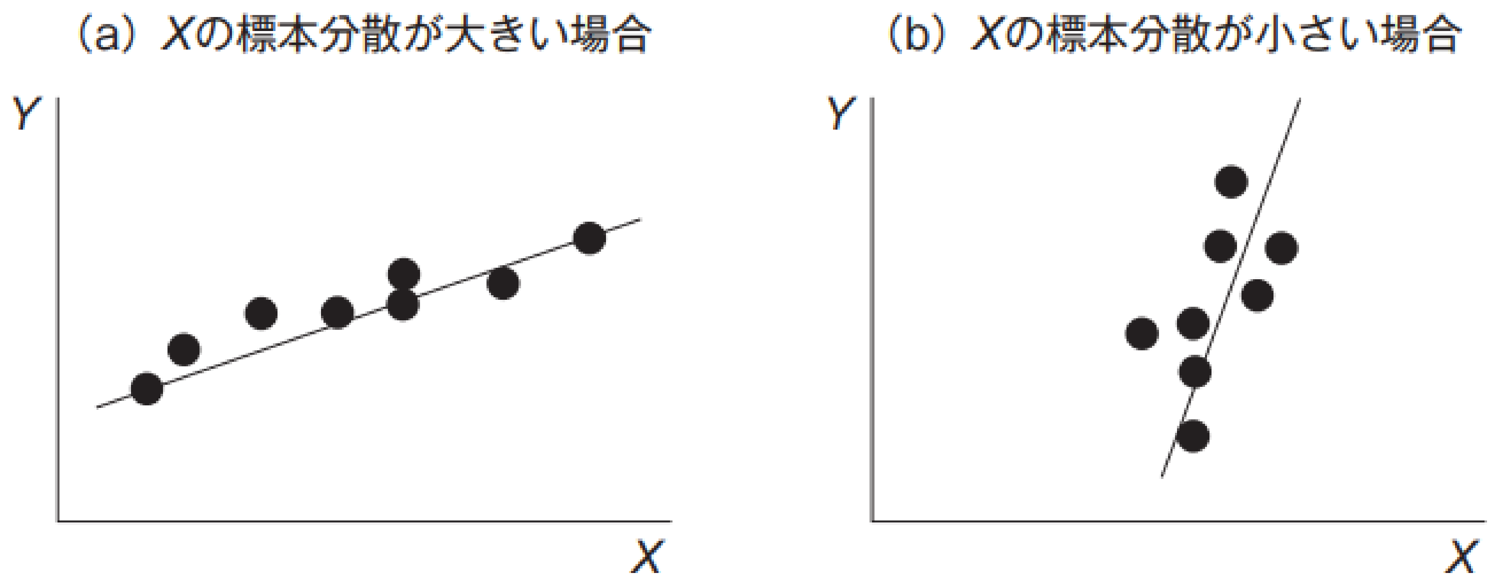
図２-６　係数と標本共分散の符号



$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})\,/(n-1)}{\boxed{\sum_{i=1}^{n}(X_i - \bar{X})^2\,/(n-1)}}$$

---Xの分散が大きい（小さい）ほど傾きはゆるい（きつい）

図 2 - 7　X の標本分散と傾きの関係



$\hat{\alpha}$の意味を考える

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

・αは、次のn個の方法でも推定できる

$$Y_1 - \hat{\beta}X_1,\ \ Y_2 - \hat{\beta}X_2, \dots,\ \ Y_n - \hat{\beta}X_n$$

[証明]
$$Y_i = \alpha + \beta X_i$$
$$\Rightarrow\ \ \alpha = Y_i - \beta X_i \quad [終]$$

・ これらn個の平均をとったものが$\hat{\alpha}$となる

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}X_i)}{n} = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}\sum_{i=1}^{n} X_i}{n}$$
$$= \underbrace{\frac{\sum_{i=1}^{n} Y_i}{n}}_{\bar{Y}} - \hat{\beta}\underbrace{\frac{\sum_{i=1}^{n} X_i}{n}}_{\bar{X}}$$

回帰直線と予測

重要な概念

- **回帰直線**: 残差2乗和を最小にする直線

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

- **予測値(理論値)**: Xに具体的値X_iを代入して求めた
  回帰直線上の値

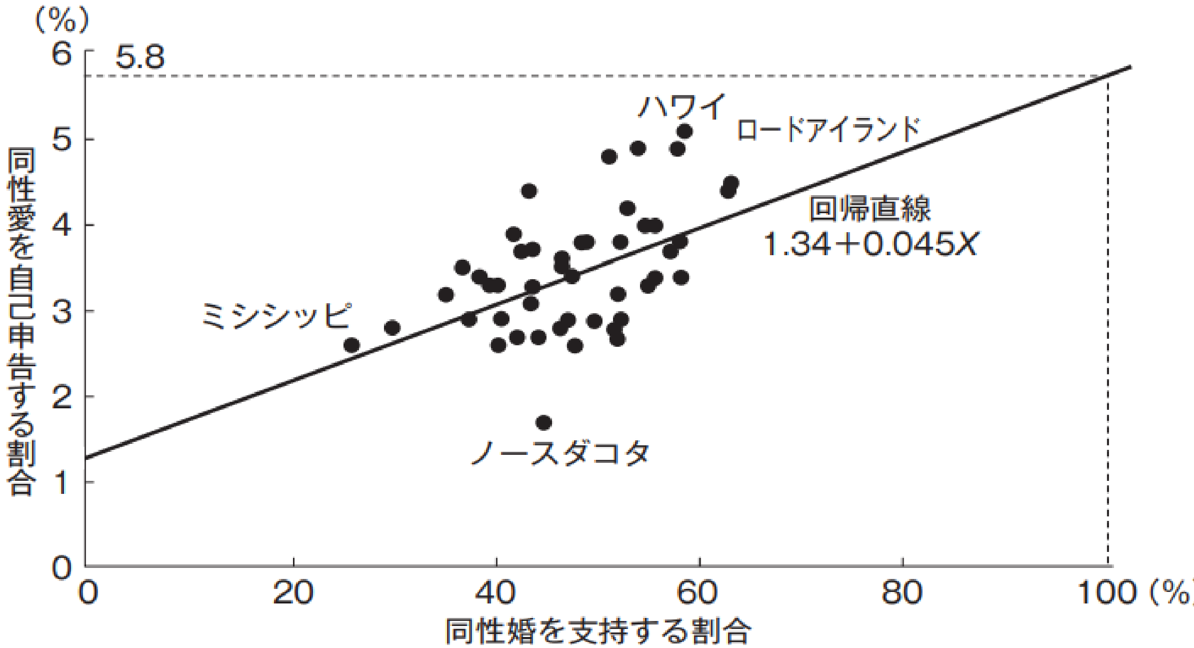
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

- **残差**: 観測値Y_iと予測値$\hat{Y}_i$との差

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$$

- **実現値の分解**: $Y_i = \hat{Y}_i + \hat{u}_i = \underbrace{\hat{\alpha} + \hat{\beta} X_i}_{\text{モデルで説明できた部分}} + \underbrace{\hat{u}_i}_{\text{モデルで説明できなかった部分}}$

例）同性愛者の割合(米国州別データ)



$$\hat{Y}_i = 1.34 + 0.045X_i$$

--- X_i : 同性婚を支持する割合、Y_i : 同性愛を自己申告する割合

- $\hat{\beta} = 0.045$: 支持割合が1%ポイント上がると、
  自己申告が0.045%ポイント増える
- $\hat{\alpha} = 1.34$は、支持割合が0%のとき、自己申告は1.34%しかない
- 支持が100%なら、自己申告の予測値は、
  $1.34 + 0.045 \times 100 = 5.8\%$

最小2乗推定量の導出

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- 残差2乗和を最小化する$\tilde{\alpha}$と$\tilde{\beta}$を求める

$$\sum_{i=1}^{n} \tilde{u}_i^2 = \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta} X_i)^2$$

- $\tilde{\alpha}$と$\tilde{\beta}$でそれぞれ偏微分して0と置く

$$\frac{\partial \sum \tilde{u}_i^2}{\partial \tilde{\alpha}} = \sum \frac{\partial \tilde{u}_i^2}{\partial \tilde{\alpha}} = \sum \frac{\partial \tilde{u}_i^2}{\partial \tilde{u}_i} \frac{\partial \tilde{u}_i}{\partial \tilde{\alpha}}$$

合成関数の微分の公式

$$= \sum 2\tilde{u}_i(-1) = \underline{-2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0}$$

$$\frac{\partial}{\partial \tilde{\beta}} \sum \tilde{u}_i^2 = \sum \frac{\partial \tilde{u}_i^2}{\partial \tilde{\beta}} = \sum \frac{\partial \tilde{u}_i^2}{\partial \tilde{u}_i} \frac{\partial \tilde{u}_i}{\partial \tilde{\beta}}$$

$$= \sum 2\tilde{u}_i(-X_i) = \underline{-2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0}$$

- 正規方程式

① $\sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$

② $\sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0$

- $\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$から、以下の残差の性質が得られる

①' $\sum \hat{u}_i = 0$ (残差の和は0)

②' $\sum \hat{u}_i X_i = 0$ (残差と説明変数の積和は0)

・正規方程式①から

$$\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i) = \sum Y_i - n\hat{\alpha} - \hat{\beta}\sum X_i = 0$$

両辺をnで割ると

$$\frac{\sum Y_i}{n} - \hat{\alpha} - \hat{\beta}\frac{\sum X_i}{n} = 0$$

となり、これを$\hat{\alpha}$について解くと$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$となる。

・正規方程式②に、$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$を代入すると

$$\begin{aligned}\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i &= \sum(Y_i - (\bar{Y} - \hat{\beta}\bar{X}) - \hat{\beta}X_i)X_i \\ &= \sum(Y_i - \bar{Y} - \hat{\beta}(X_i - \bar{X}))X_i \\ &= \sum(Y_i - \bar{Y})X_i - \hat{\beta}\sum\left(X_i - \bar{X}\right)X_i = 0\end{aligned}$$

よって、

$$\hat{\beta} = \frac{\sum(Y_i - \bar{Y})X_i}{\sum\left(X_i - \bar{X}\right)X_i} = \frac{\sum(Y_i - \bar{Y})\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2}$$

偏差の和は0

- $\sum\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) = \sum\left(Y_i - \bar{Y}\right)X_i - \bar{X}\sum\left(Y_i - \bar{Y}\right) = \sum\left(Y_i - \bar{Y}\right)X_i$

$$\sum\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right) = \sum\left(X_i - \bar{X}\right)X_i - \bar{X}\sum\left(X_i - \bar{X}\right) = \sum\left(X_i - \bar{X}\right)X_i$$

・偏差の和は0

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n} Y_i - n\bar{Y}$$

$$= \sum_{i=1}^{n} Y_i - n\frac{\sum_{i=1}^{n} Y_i}{n}$$

$$= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Y_i = 0$$

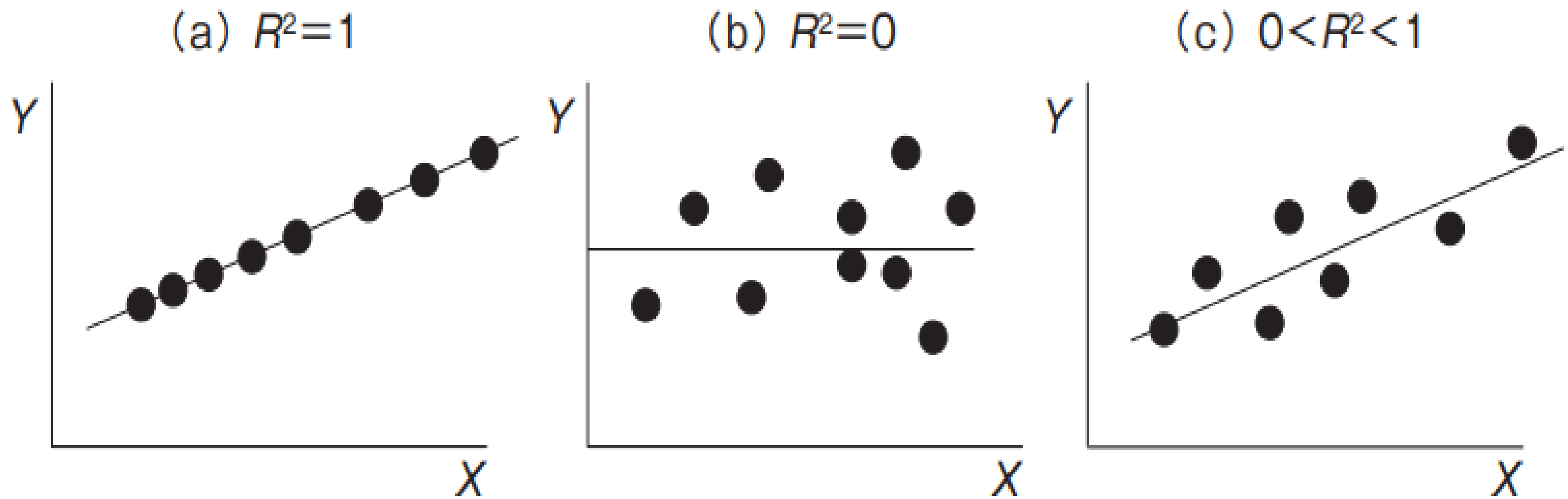
--- 数式展開で用いられる性質

決定係数

決定係数：　当てはまりの良さを測る指標

--- 0〜1で定義、1に近いほど当てはまり良い

図 2-10　決定係数



--- Yの変動をどれぐらい予測値の変動で説明できるか

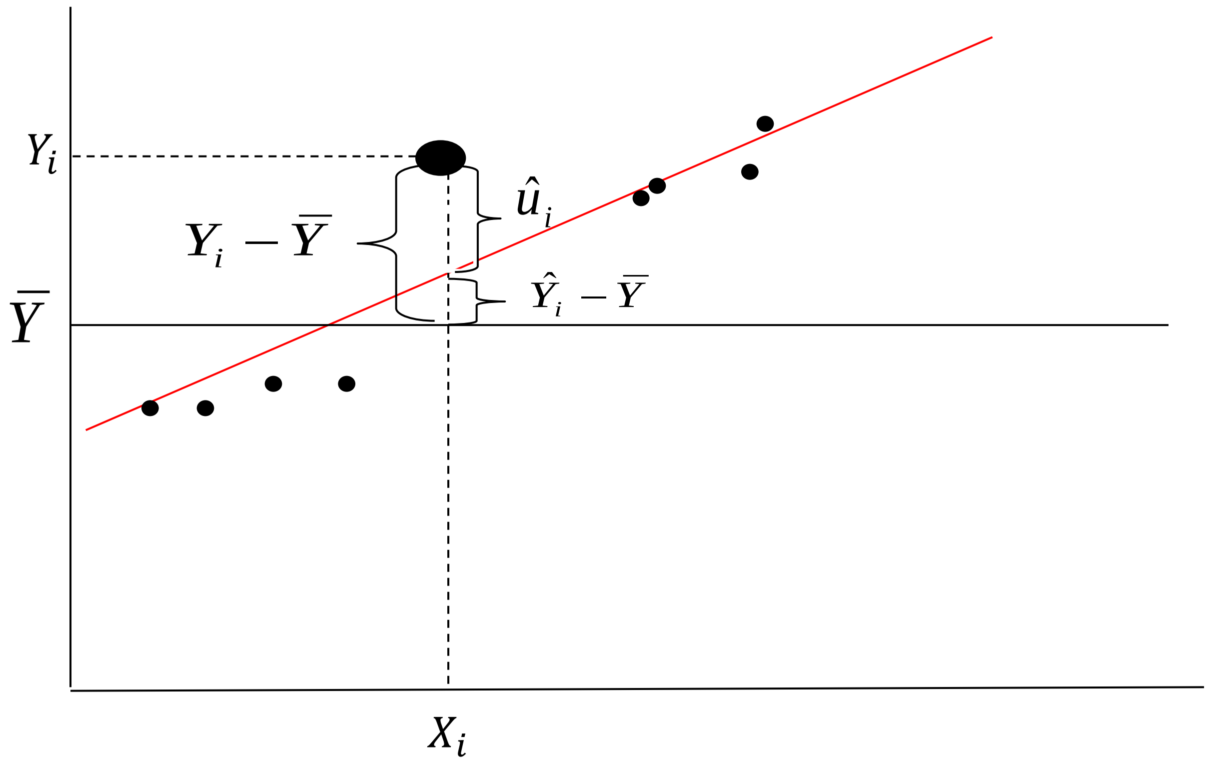
--- 実現値の分解

$$Y_i = \hat{Y}_i + \hat{u}_i = \underbrace{\hat{\alpha} + \hat{\beta} X_i}_{\text{モデルで説明できた部分}} + \underbrace{\hat{u}_i}_{\text{モデルで説明できなかった部分}}$$

- Yの全変動を分解する

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{Y\text{の全変動}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{モデルで説明された}Y\text{の変動}} + \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{\text{モデルで説明されなかった}Y\text{の変動}}$$

図２-９　偏差の分解



[証明]

・ここで$Y_i = \hat{Y}_i + \hat{u}_i$から、$Y_i$の全変動は

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i + \hat{u}_i - \bar{Y})^2$$
$$= \sum_{i=1}^{n}((\hat{Y}_i - \bar{Y}) + \hat{u}_i)^2$$
$$= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{u}_i^2 + 2 \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\,\hat{u}_i}_{=0}$$

[第3項目が0の証明]

残差の性質は、

① $\sum \hat{u}_i = 0$ (残差の和は0)

② $\sum \hat{u}_i X_i = 0$ (残差と説明変数の積和は0)

①②を用いると、

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\,\hat{u}_i = \sum_{i=1}^{n} \hat{Y}_i\,\hat{u}_i - \bar{Y}\sum_{i=1}^{n} \hat{u}_i = 0$$

なお、

$$\sum_{i=1}^{n} \hat{Y}_i\,\hat{u}_i = \sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta} X_i)\,\hat{u}_i$$
$$= \hat{\alpha}\sum_{i=1}^{n} \hat{u}_i + \hat{\beta}\sum_{i=1}^{n} X_i\,\hat{u}_i = 0$$

- **決定係数:モデルで説明された変動÷Yの全変動**

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

$$=1-\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

式展開

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}\hat{u}_i^2$$

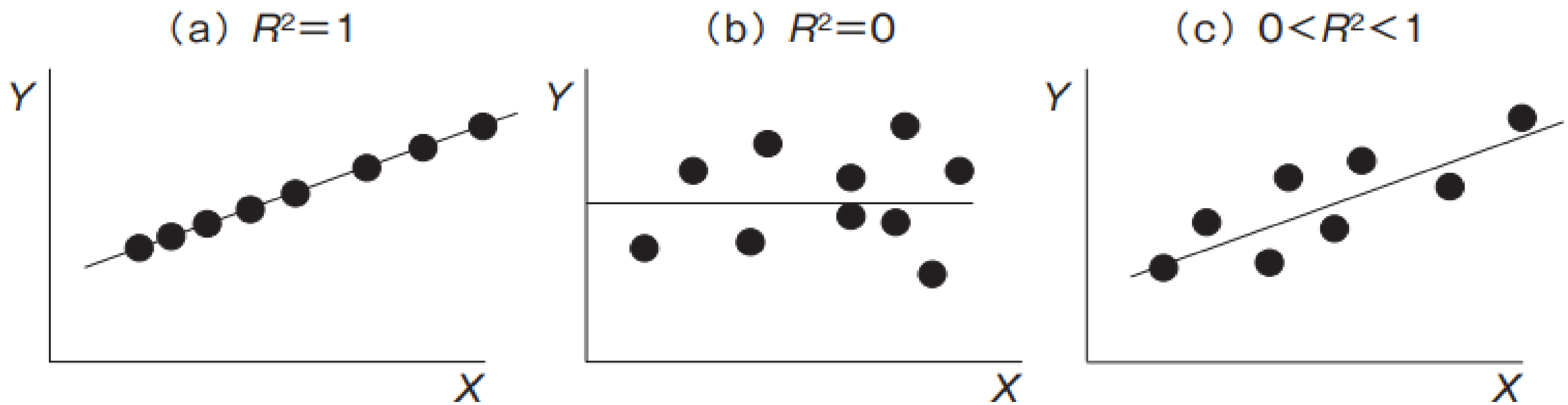
$$\Rightarrow \quad \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\sum_{i=1}^{n}\hat{u}_i{}^2$$

よって、

$$\frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}=\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\sum_{i=1}^{n}\hat{u}_i{}^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

$$=1-\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

図 2-10　決定係数



(a) 残差は0なのでR^2は1となる

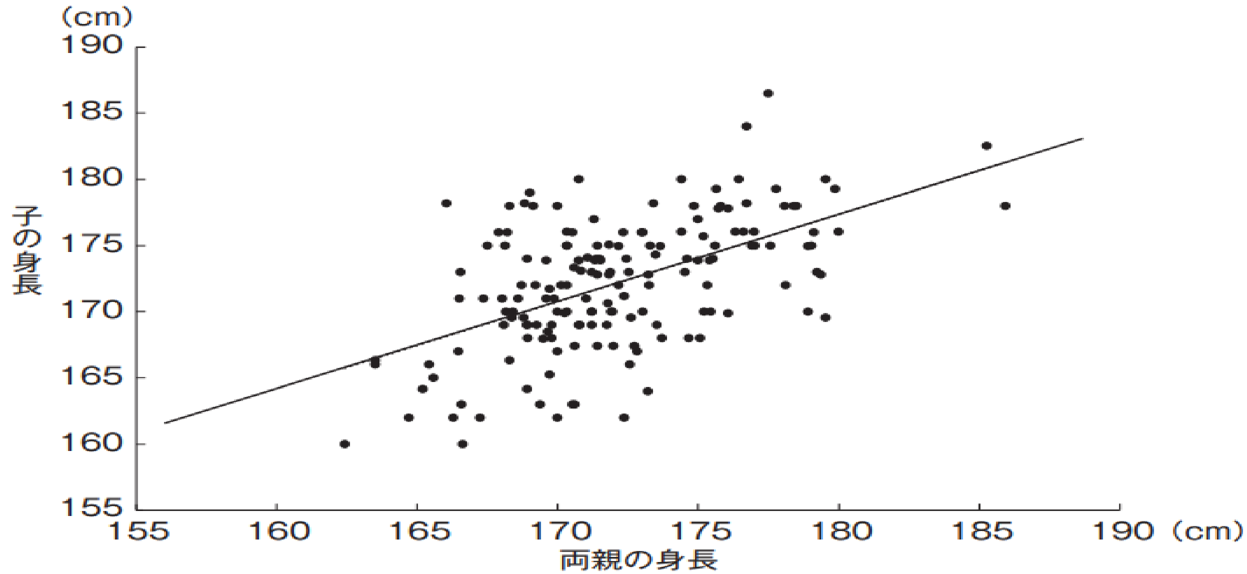
(b) 残差が大きいためR^2は0に近い値となる

(c) 通常、R^2は0より大きく、1より小さい

例）両親と子供の身長

- 両親の平均身長 $=$ (父親の身長 $+ 1.08 \times$ 母親の身長$)/2$
- 子供の身長、ただし、娘の身長は$1.08 \times$ 娘の本当の身長

図 2 -11　親子の身長の関係



$$\widehat{\text{子供の身長}}_i = 58.41 + 0.661\text{両親の身長}_i$$

- $\beta = 0.661$: 両親が1cm高いと子供の身長も0.661cm高くなる
- $\alpha = 58.41$は数学的切片と解釈する
- 決定係数は0.295 (子供の身長の変動のうち29.5%が親の身長から説明できる)

- 両親の平均身長190cmなら、子供は

$$58.41 + 0.661 \times 190 = 184cm$$

両親の平均身長160cmなら、子供は

$$58.41 + 0.661 \times 160 = 164cm$$

ゴールトンの「平均への回帰」

親の身長が高いと子供の身長も高くなるが、

親の身長ほど高くはない

まとめ

- **回帰モデル**

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

- **最小2乗法**

  --- 当てはまりの良い線を引く

- **予測**

  --- 推定された回帰直線に基づいた予測

- **決定係数**

  --- 当てはまりの尺度(0～1で定義される)