第4章　回帰分析

4.1　準備

4.1.1　重要な公式

1. $\displaystyle\sum_{i=1}^{n} X_i = n\overline{X}$

2. $\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}) = 0$

3. $\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$

4. $\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} = \sum_{i=1}^{n}(X_i - \overline{X})Y_i = \sum_{i=1}^{n}(Y_i - \overline{Y})X_i$

5. 2×2 行列の逆行列の公式： $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \dfrac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

4.1.2　データについて

1. タイム・シリーズ (時系列)・データ： 添え字 i が時間を表す (第 i 期)。t を添え字に使う場合も多い。

2. クロス・セクション (横断面)・データ： 添え字 i が個人や企業を表す (第 i 番目の家計，第 i 番目の企業)。

4.2　最小二乗法について：単回帰モデル

最小二乗法とは，線型モデルの係数の値をデータから求める時に用いられる手法である。

4.2.1　最小二乗法と回帰直線

$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ のように n 組のデータがあり，X_i と Y_i との間に以下の線型関係を想定する。

$$Y_i = \alpha + \beta X_i,$$

X_i は説明変数，Y_i は被説明変数，α, β はパラメータとそれぞれ呼ばれる。

上の式は回帰モデル（または，回帰式）と呼ばれる。切片 α と傾き β をデータ $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ から求める（推定する）ことを考える。

ある基準の下で，α と β の推定値が求められたとしよう。それぞれ，$\hat{\alpha}$ と $\hat{\beta}$ とする。データ $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ と直線との関係は,

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i,$$

となる。すなわち，実際のデータ Y_i と直線上の値 $\hat{\alpha} + \hat{\beta}X_i$ との間には，差 $\hat{u}_i$（残差と呼ばれる）が生じる。

4.2.2　切片 α と傾き β の求め方

α, β のある推定値を $\hat{\alpha}, \hat{\beta}$ としよう。次のような関数 $S(\hat{\alpha}, \hat{\beta})$ を定義する。

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

この $S(\hat{\alpha}, \hat{\beta})$ は残差平方和と呼ばれる。

このとき，

$$\min_{\hat{\alpha},\, \hat{\beta}} S(\hat{\alpha}, \hat{\beta})$$

となるような $\hat{\alpha}, \hat{\beta}$ を求める（最小自乗法）。

最小化のためには,

$$\frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = 0, \qquad \frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = 0$$

を満たす $\hat{\alpha}, \hat{\beta}$ を求める。

　すなわち，$\hat{\alpha}, \hat{\beta}$ は,

$$\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0, \tag{4.1}$$

$$\sum_{i=1}^{n} X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0, \tag{4.2}$$

を満たす。

　さらに,

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i \tag{4.3}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2 \tag{4.4}$$

(4.3) 式の辺々を n で割って,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \hat{\alpha} + \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} X_i$$

すなわち，

$$\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X} \tag{4.5}$$

を得る。ただし，

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

とする。

さらに，$\sum_{i=1}^{n} X_i = n\overline{X}$ と (4.5) 式を利用して，$\hat{\alpha}$ を消去すると，

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

$\hat{\beta}$ で整理して，

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{S_{XY}}{S_X^2} \tag{4.6}$$

が得られ，$\hat{\alpha}$ は (4.5) 式から，

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \tag{4.7}$$

となる。ただし，

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}), \qquad S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

とする。

または，行列を用いて解くこともできる。行列表示によって，

$$\begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix},$$

$\hat{\alpha}$，$\hat{\beta}$ について，まとめて，

$$\begin{aligned} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} &= \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix} \end{aligned}$$

さらに，$\hat{\beta}$ について解くと，

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

$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

$\hat{\alpha}$ については,

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

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

$$= \overline{Y} - \hat{\beta}\overline{X}$$

となる。

回帰直線は,

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

として与えられる。$\hat{Y}_i$ は，X_i を与えたときの Y_i の予測値と解釈される。

数値例：　以下の数値例を使って，回帰式 $Y_i = \alpha + \beta X_i$ の α，β の推定値 $\hat{\alpha}$，$\hat{\beta}$ を求める。

i	X_i	Y_i
1	5	4
2	1	1
3	3	1
4	2	3
5	4	4

$\hat{\alpha}$，$\hat{\beta}$ を求めるための公式は，

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}, \qquad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X},$$

なので，必要なものは $\overline{X}$，$\overline{Y}$，$\displaystyle\sum_{i=1}^{n} X_i^2$，$\displaystyle\sum_{i=1}^{n} X_i Y_i$ である。

i	X_i	Y_i	X_i^2	X_iY_i
1	5	4	25	20
2	1	1	1	1
3	3	1	9	3
4	2	3	4	6
5	4	4	16	16
合計	$\sum X_i$	$\sum Y_i$	$\sum X_i^2$	$\sum X_iY_i$
	15	13	55	46
平均	$\overline{X}$	$\overline{Y}$		
	3	2.6		

表中では，$\sum_{i=1}^{n}$ を $\sum$ と省略して表記している。

よって，

$$\hat{\beta} = \frac{46 - 5 \times 3 \times 2.6}{55 - 5 \times 3^2} = \frac{7}{10} = 0.7, \qquad \hat{\alpha} = 2.6 - 0.7 \times 3 = 0.5,$$

となる。

注意事項：

1. α, β は真の値で未知である。

2. $\hat{\alpha}, \hat{\beta}$ は α, β の推定値でデータから計算される。

回帰直線は，$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ であり，上の数値例では,

$$\hat{Y}_i = 0.5 + 0.7X_i,$$

となる。$\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_5$ として，次の表のように計算される。Y_i，X_i，$\hat{Y}_i$，$\hat{u}_i$ の関係が図 4.1 に描かれている。

i	X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$
1	5	4	25	20	4.0
2	1	1	1	1	1.2
3	3	1	9	3	2.6
4	2	3	4	6	1.9
5	4	4	16	16	3.3
合計	$\sum X_i$	$\sum Y_i$	$\sum X_i^2$	$\sum X_iY_i$	$\sum \hat{Y}_i$
	15	13	55	46	13
平均	$\overline{X}$	$\overline{Y}$			
	3	2.6			

$\hat{Y}_i$ を実績値 Y_i の予測値または理論値と呼ぶ。

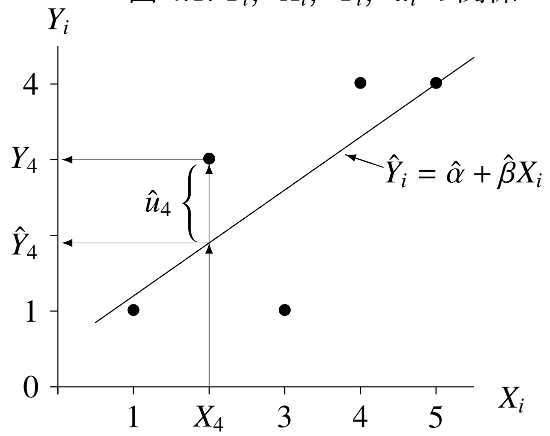
$$\hat{u}_i = Y_i - \hat{Y}_i,$$

$\hat{u}_i$ を残差と呼ぶ。$Y_i, \hat{Y}_i, \hat{u}_i$ の関係，$\hat{Y}_i, X_i, \hat{\alpha}, \hat{\beta}$ の関係は,

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

の式でまとめられる。

図 4.1: Y_i，X_i，$\hat{Y}_i$，$\hat{u}_i$ の関係



4.2.3　残差 $\hat{u}_i$ の性質について

$\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$ に注意すると，(4.1) 式，(4.2) 式から，

$$\sum_{i=1}^{n} \hat{u}_i = 0, \qquad \sum_{i=1}^{n} X_i\hat{u}_i = 0,$$

を得る。また，$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ から，

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

が得られる。なぜなら，

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

となるからである。

数値例で確認してみよう。

i	X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$	$\hat{u}_i$	$X_i\hat{u}_i$	$\hat{Y}_i\hat{u}_i$
1	5	4	25	20	4.0	0.0	0.0	0.00
2	1	1	1	1	1.2	−0.2	−0.2	−0.24
3	3	1	9	3	2.6	−1.6	−4.8	−4.16
4	2	3	4	6	1.9	1.1	2.2	2.09
5	4	4	16	16	3.3	0.7	2.8	2.31
合計	$\sum X_i$	$\sum Y_i$	$\sum X_i^2$	$\sum X_iY_i$	$\sum \hat{Y}_i$	$\sum \hat{u}_i$	$\sum X_i\hat{u}_i$	$\sum \hat{Y}_i\hat{u}_i$
	15	13	55	46	13	0.0	0.0	0.0
平均	$\overline{X}$	$\overline{Y}$						
	3	2.6						

4.2.4　決定係数 R^2 について

Y_i, $\hat{Y}_i$, $\hat{u}_i$ の関係は,

$$Y_i = \hat{Y}_i + \hat{u}_i,$$

であった。$\overline{Y}$ を両辺から引くと,

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + \hat{u}_i,$$

が得られる。さらに，両辺を二乗して，総和すると,

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \sum_{i=1}^{n}((\hat{Y}_i - \overline{Y}) + \hat{u}_i)^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})\hat{u}_i + \sum_{i=1}^{n}\hat{u}_i^2 \\ &= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}\hat{u}_i^2\end{aligned}$$

となる。二つ目の等式の右辺第二項では，$\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i = \overline{Y}\sum_{i=1}^{n}\hat{u}_i = 0$ が使われている。まとめると,

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

を得る。さらに，両辺を左辺で割ると,

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

が得られる。それぞれの項は,

1. $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \longrightarrow Y_i$ の全変動

2. $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \longrightarrow \hat{Y}_i$ (回帰直線) で説明される部分

3. $\sum_{i=1}^{n}\hat{u}_i^2 \longrightarrow \hat{Y}_i$ (回帰直線) で説明されない部分

となる。

回帰式の当てはまりの良さを示す指標として，決定係数 R^2 が，

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}, \tag{4.8}$$

のように定義される。R^2 は Y_i のうち $\hat{Y}_i$（または，X_i）で説明できる比率を意味する。または，

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}, \tag{4.9}$$

として書き換えることもできる。

R^2 の取り得る範囲:　さらに，R^2 の取り得る範囲を求める。(4.8) 式の右辺の分子と分母は共に正なので，$R^2 \geq 0$ となる。(4.9) 式の右辺では 1 から第二項の正の値（分子分母共に正）を差し引いているので，$R^2 \leq 1$ となることが分かる。すなわち，R^2 の取り得る範囲は,

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

となる。

$R^2 = 1$ となる場合はすべての i について $\hat{u}_i = 0$ となり，観測されたデータ (X_i, Y_i) は一直線上に並んでいる状態となる。

$R^2 = 0$ となる場合は二通りが考えられる。一つは，Y_i が X_i に影響されないときで，$\hat{\beta} = 0$ の状態，すなわち，データが横軸に平行に一直線上に並んでいる状態となる。もう一つは，データが円状に散布していて，どこにも直線が引けない状態である（ちなみに，データが楕円上に散布している場合は，直線が引ける状態である）。

実際のデータを用いた場合は $R^2 = 0$ や $R^2 = 1$ という状況はあり得ない。R^2 が 1 に近づけば回帰式の当てはまりは良い，R^2 が 0 に近づけば回帰式の当てはまりは悪いと言える。しかし,「どの値よりも大きくなるべき」といった基準はない。慣習的には，メドとして 0.9 以上が当てはまりが良いと判断する。

データと R^2 との関係は，後述の 4.2.5 節で，数値例を挙げながら解説する。

R^2 の別の解釈:　R^2 のもう一つの解釈をするために，R^2 の右辺の分子を，

$$\begin{aligned}\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})(Y_i-\overline{Y}-\hat{u}_i) = \sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})(Y_i-\overline{Y}) - \sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})\hat{u}_i \\ &= \sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})(Y_i-\overline{Y}),\end{aligned}$$

と書き換える。最初の等式では，括弧二乗の一つに $\hat{Y}_i = Y_i - \hat{u}_i$ が用いられている。R^2 は，

$$\begin{aligned}R^2 &= \frac{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2} = \frac{\left(\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2\right)^2}{\left(\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)\left(\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2\right)} \\ &= \left(\frac{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}}\right)^2,\end{aligned}$$

と書き換えられる。この式では，R^2 が Y_i と $\hat{Y}_i$ の相関係数の二乗と解釈されることを意味する。なお，二つ目の等号の右式では，分子と分母に $\sum_{i=1}^n(\hat{Y}_i - \overline{Y})^2$ を掛けていることに注意せよ。

特に，単回帰の場合，$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ と $\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X}$ を用いて，

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

を利用すると，

$$R^2 = \frac{\sum_{i=1}^n(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n(Y_i - \overline{Y})^2} = \frac{\hat{\beta}^2\sum_{i=1}^n(X_i - \overline{X})^2}{\sum_{i=1}^n(Y_i - \overline{Y})^2} = \left(\frac{\sum_{i=1}^n(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n(Y_i - \overline{Y})^2}\sqrt{\sum_{i=1}^n(X_i - \overline{X})^2}}\right)^2$$

$$= \frac{S_{XY}^2}{S_X^2 S_Y^2},$$

としても書き換えられる。すなわち，単回帰の場合，決定係数は説明変数 X_i と被説明変数 Y_i との相関係数の二乗となる。

数値例：　決定係数の計算には以下の公式を用いる。

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2}$$

計算に必要なものは，$\sum_{i=1}^{n} \hat{u}_i^2$，$\overline{Y}$，$\displaystyle\sum_{i=1}^{n} Y_i^2$ である。

i	X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$	$\hat{u}_i$	$X_i\hat{u}_i$	$\hat{Y}_i\hat{u}_i$	$\hat{u}_i^2$	Y_i^2
1	5	4	25	20	4.0	0.0	0.0	0.00	0.00	16
2	1	1	1	1	1.2	−0.2	−0.2	−0.24	0.04	1
3	3	1	9	3	2.6	−1.6	−4.8	−4.16	2.56	1
4	2	3	4	6	1.9	1.1	2.2	2.09	1.21	9
5	4	4	16	16	3.3	0.7	2.8	2.31	0.49	16
合計	$\sum X_i$	$\sum Y_i$	$\sum X_i^2$	$\sum X_iY_i$	$\sum \hat{Y}_i$	$\sum \hat{u}_i$	$\sum X_i\hat{u}_i$	$\sum \hat{Y}_i\hat{u}_i$	$\sum \hat{u}_i^2$	$\sum Y_i^2$
	15	13	55	46	13	0.0	0.0	0.0	4.3	43
平均	$\overline{X}$	$\overline{Y}$								
	3	2.6								

$\overline{Y} = 2.6,\ \sum_{i=1}^{n} \hat{u}_i^2 = 4.3,\ \sum_{i=1}^{n} Y_i^2 = 43$ なので,

$$R^2 = 1 - \frac{4.3}{43 - 5 \times 2.6^2} = \frac{4.9}{9.2} = 0.5326$$

4.2.5　決定係数の比較

次の数値例を用いて，決定係数の比較を行おう。X と Y のプロットしたものが図 4.2(a)~(d) である。

	(a)		(b)		(c)		(d)	
i	X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i
1	1	1	1	1	1	1.5	1	3
2	2	1	2	1.5	2	2.3	2.5	2.134
3	2	3	2	2.5	3	3.1	2.5	3.866
4	4	3	4	3.5	3.5	3.5	3.5	2.134
5	4	5	4	4.5	4	3.9	3.5	3.866
6	5	5	5	5	5	4.7	4	3

(a) と (b) のどちらの場合も，切片・傾きの値は $\hat{\alpha} = 0$，$\hat{\beta} = 1$ として計算されるが，決定係数について，(a) は 0.75，(b) は 0.923 となる（読者はチェックすること）。データのプロットと回帰直線は図 4.2 の (a) と (b) に描かれている。X_i はどちらも同じ数値とした。横軸 X が 2，4 のケースについて，(b) が (a) より直線に近くなるように，Y の値を変えてみた。(b) のデータの方が (a) より直線に近いために，決定係数が 0.923 と 1 に近い値となっているのが分かる。

(c) はデータが一直線上に並んでいる場合で，決定係数が 1 となる。決定係数がゼロとなるのは (d) の場合で，X と Y との関係を表す直線が描けない場合である。(d) の数値例では，X と Y との関係が円としているが，満遍なく散布している状態と考えてもらえれば良い。

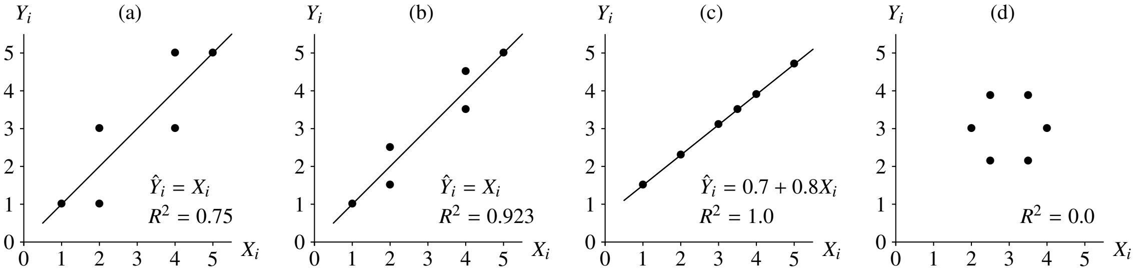
4.2.6　まとめ

$\hat{\alpha}$，$\hat{\beta}$ を求めるための公式は

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

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

図 4.2: 決定係数の比較



なので，必要なものは $\overline{X}$，$\overline{Y}$，$\displaystyle\sum_{i=1}^{n} X_i^2$，$\displaystyle\sum_{i=1}^{n} X_i Y_i$ である。

決定係数の計算には以下の公式を用いる。

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2}$$

ただし，$\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$ である。計算に必要なものは，$\sum_{i=1}^{n} \hat{u}_i^2$，$\overline{Y}$，$\displaystyle\sum_{i=1}^{n} Y_i^2$ で

ある。

4.3　最小二乗法について：重回帰モデル

k 変数の多重回帰モデルを考える。

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

X_{ji} は j 番目の説明変数の第 i 番目の観測値を表す。$\beta_1, \beta_2, \cdots, \beta_k$ は推定されるべきパラメータである。すべての i について，$X_{1i} = 1$ とすれば，β_1 は定数項として表される。n 組のデータ $(Y_i, X_{1i}, X_{2i}, \cdots, X_{ki})$, $i = 1, 2, \cdots, n$ を用いて，$\beta_1, \beta_2, \cdots, \beta_k$ を求める。

ある基準の下で，$\beta_1, \beta_2, \cdots, \beta_k$ の解を $\hat\beta_1, \hat\beta_2, \cdots, \hat\beta_k$ としよう。データ $\{(X_i, Y_i),\ i = 1, 2, \cdots, n\}$ と直線との関係は，

$$Y_i = \hat\beta_1 X_{1i} + \hat\beta_2 X_{2i} + \cdots + \hat\beta_k X_{ki} + \hat u_i = \hat Y_i + \hat u_i,$$

となる。すなわち，すべての i について，実際のデータ Y_i と直線上の値 $\hat Y_i = \hat\beta_1 X_{1i} + \hat\beta_2 X_{2i} + \cdots + \hat\beta_k X_{ki}$ が一致することはあり得ないので，残差 $\hat u_i$ の二乗和を考える。

　次のような関数 $S(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)$ を定義する。

$$S(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2$$

このとき,

$$\min_{\hat{\beta}_1,\, \hat{\beta}_2,\, \cdots,\, \hat{\beta}_k} S(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)$$

となるような $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ を求める。$\Longrightarrow$ 最小自乗法

　最小化のためには,

$$\frac{\partial S(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)}{\partial \hat{\beta}_1} = 0, \quad \frac{\partial S(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)}{\partial \hat{\beta}_2} = 0, \quad \cdots, \quad \frac{\partial S(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)}{\partial \hat{\beta}_k} = 0$$

を満たす $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ となる。

　すなわち，$\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ は,

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}) X_{1i} = 0,$$

$$\sum_{i=1}^{n}(Y_i - \hat\beta_1 X_{1i} - \hat\beta_2 X_{2i} - \cdots - \hat\beta_k X_{ki})X_{2i} = 0,$$
$$\vdots$$
$$\sum_{i=1}^{n}(Y_i - \hat\beta_1 X_{1i} - \hat\beta_2 X_{2i} - \cdots - \hat\beta_k X_{ki})X_{ki} = 0,$$

を満たす。

さらに,

$$\sum_{i=1}^{n} X_{1i}Y_i = \hat\beta_1 \sum_{i=1}^{n} X_{1i}^2 + \hat\beta_2 \sum_{i=1}^{n} X_{1i}X_{2i} + \ \cdots \ + \hat\beta_k \sum_{i=1}^{n} X_{1i}X_{ki},$$
$$\sum_{i=1}^{n} X_{2i}Y_i = \hat\beta_1 \sum_{i=1}^{n} X_{1i}X_{2i} + \hat\beta_2 \sum_{i=1}^{n} X_{2i}^2 + \ \cdots \ + \hat\beta_k \sum_{i=1}^{n} X_{2i}X_{ki},$$
$$\vdots$$
$$\sum_{i=1}^{n} X_{ki}Y_i = \hat\beta_1 \sum_{i=1}^{n} X_{1i}X_{ki} + \hat\beta_2 \sum_{i=1}^{n} X_{2i}X_{ki} + \ \cdots \ + \hat\beta_k \sum_{i=1}^{n} X_{ki}^2,$$

行列表示によって,

$$\begin{pmatrix} \sum X_{1i}Y_i \\ \sum X_{2i}Y_i \\ \vdots \\ \sum X_{ki}Y_i \end{pmatrix} = \begin{pmatrix} \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \cdots & \sum X_{1i}X_{ki} \\ \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \cdots & \sum X_{2i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \cdots & \sum X_{ki}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

が得られ, $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ についてまとめると,

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \cdots & \sum X_{1i}X_{ki} \\ \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \cdots & \sum X_{2i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \cdots & \sum X_{ki}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum X_{1i}Y_i \\ \sum X_{2i}Y_i \\ \vdots \\ \sum X_{ki}Y_i \end{pmatrix}$$

を解くことになる。$\Longrightarrow$ コンピュータによって計算

4.3.1　重回帰モデルにおける回帰係数の意味

結論： 他の変数の影響を取り除いての被説明変数への影響を表す。