3.2.4　決定係数 R^2 について

3.2.2 節では，(X_1, Y_1)，(X_2, Y_2)，$\cdots$，(X_n, Y_n) のデータから，X と Y との関係を表す直線の引き方を学んだ。

データさえあれば直線を引くことは可能であるが，本当に X と Y との関係が直線で表せるのかどうかを調べる必要がある。

データと直線との当てはまりの良さを示す指標として，決定係数が用いられ，R^2 という記号で表される。

まず準備として，$Y_i, \hat{Y}_i, \hat{u}_i$ の関係は，

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

を思い出してもらう。

$\overline{Y}$ 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。

両辺を二乗して，$(Y_i - \overline{Y})^2 = \left((\hat{Y}_i - \overline{Y}) + \hat{u}_i\right)^2$

さらに，総和して，$\sum_i^n (Y_i - \overline{Y})^2 = \sum_i^n \left((\hat{Y}_i - \overline{Y}) + \hat{u}_i\right)^2$ **を計算する。**

すなわち，

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

が得られる。

3 つ目の等式では，$\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i=\overline{Y}\sum_{i=1}^{n}\hat{u}_i=0$ (3.2.3 節参照) 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，

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}, \tag{3.8}$$

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，

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}, \tag{3.9}$$

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。

(3.8) 式の右辺の分子と分母は共に正なので，$R^2 \geq 0$ となる。

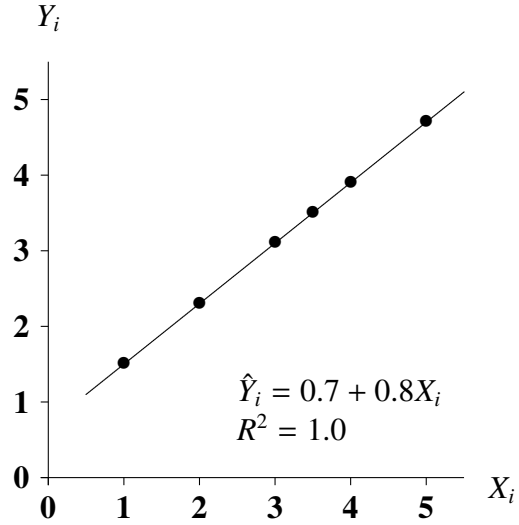
(3.9) 式の右辺では 1 から第二項の正の値（分子分母共に正）を差し引いているので，$R^2 \leq 1$ となることが分かる。

すなわち，R^2 の取り得る範囲は，

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

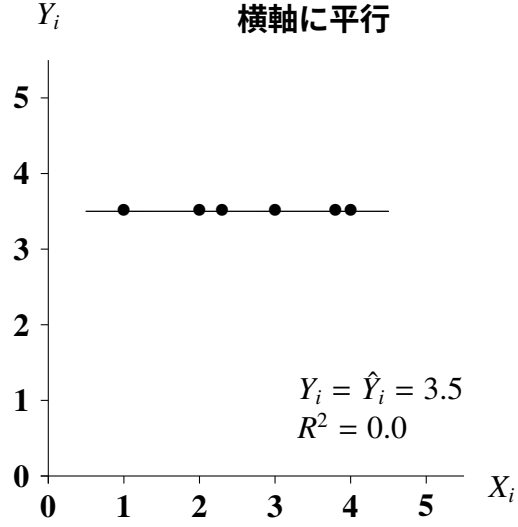
となる。

$R^2 = 1$ 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（例外は後述）。

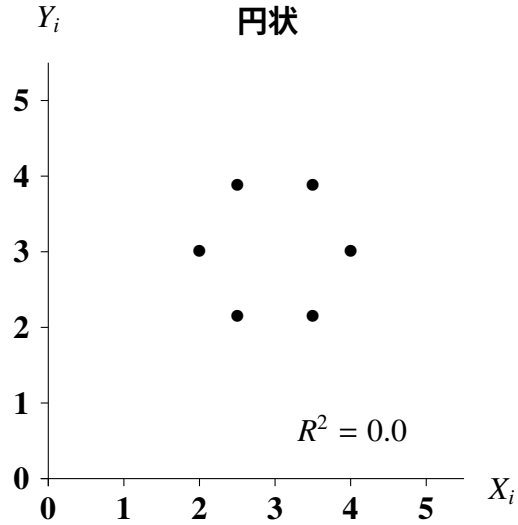


$R^2 = 0$ となる場合は二通りが考えられる。

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（すべての i について，残差はゼロではあるが ••• ）。



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 3.2.5 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，**

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

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。**

2 つ目の等式では， $\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i = 0$**，** $\sum_{i=1}^{n}\hat{u}_i = 0$ **が用いられている。**

R^2 は，次のように書き換えられる。

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

この式では，R^2 は 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，2 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。

(*) X と Y の標本相関係数は下記の通り。

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2} \sqrt{\sum (Y_i - \overline{Y})^2}}$$

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，両辺の辺々の差を取る。

$$(\hat{Y}_i - \overline{Y}) = \hat{\beta}(X_i - \overline{X})$$

さらに，両辺の二乗を取って，i を 1 から n まで足し合わせと，

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

となる。

2 つ目の等式では，$\hat{\beta} = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$ が代入されている。

したがって，$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ を書き換えると，

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

が得られる。

ただし，$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$，$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$，$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ である。

すなわち，単回帰の場合（すなわち，説明変数が X_i だけ場合），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。

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

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

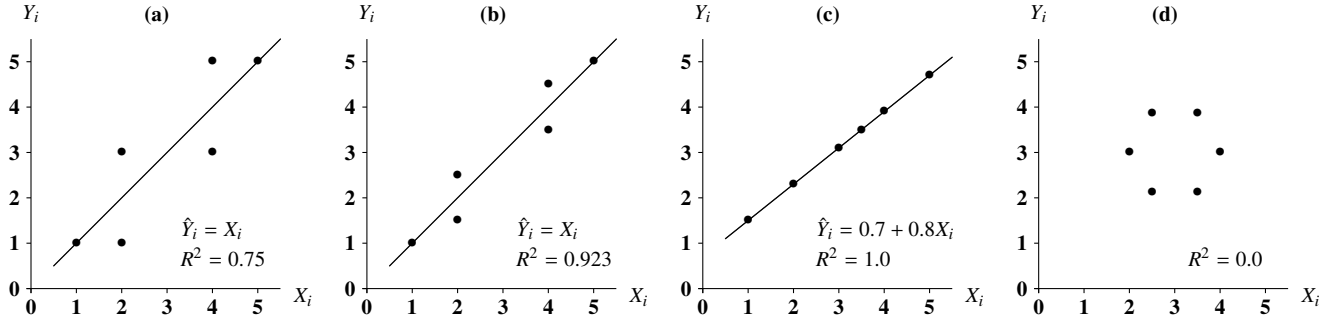
$\overline{Y} = 2.6$，$\sum_{i=1}^{n} \hat{u}_i^2 = 4.3$，$\sum_{i=1}^{n} Y_i^2 = 43$ なので，

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

3.2.5　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 3.2(a)~(d) である。

図 3.2: 決定係数の比較



	(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 3.2 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，傾きがゼロでなく，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。

3.2.6　まとめ

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

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

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。

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。

3.3　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 $\{(Y_i, X_{1i}, X_{2i}, \cdots, X_{ki}), i = 1, 2, \cdots, n\}$ 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，**

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

を満たす。

さらに，

$$
\begin{aligned}
\sum_{i=1}^{n} X_{1i}Y_i &= \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{1i}X_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{1i}X_{ki}, \\
\sum_{i=1}^{n} X_{2i}Y_i &= \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{2i}X_{ki}, \\
&\vdots \\
\sum_{i=1}^{n} X_{ki}Y_i &= \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}X_{ki} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}X_{ki} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2,
\end{aligned}
$$

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