第3章　回帰分析

3.1　準備

3.1.1　重要な公式

前章と重なる部分はあるが，重要な公式であるので，再度，下記に記しておく。

1. $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ **から，** $\sum_{i=1}^{n} X_i = n\overline{X}$ **となる。**

2. $\sum_{i=1}^{n} X_i = n\overline{X} = \sum_{i=1}^{n} \overline{X}$ **なので，** $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$ **となる。**

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

$$= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

と変形できる。

3 つ目の等式では， $\sum_{i=1}^{n} X_i = n\overline{X}$ **が使われている。**

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

$\displaystyle=\sum_{i=1}^{n}X_iY_i-n\overline{XY}-n\overline{YX}+n\overline{XY}=\sum_{i=1}^{n}X_iY_i-n\overline{X}\,\overline{Y}$

3 つ目の等式では， $\displaystyle\sum_{i=1}^{n}X_i=n\overline{X}$**，** $\displaystyle\sum_{i=1}^{n}Y_i=n\overline{Y}$ **が使われている。**

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

2 つ目の等式で， $\displaystyle\sum_{i=1}^{n}(X_i-\overline{X})=\sum_{i=1}^{n}X_i-n\overline{X}=0$ **が使われている。**

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

2 つ目の等式で， $\displaystyle\sum_{i=1}^{n}(Y_i-\overline{Y})=\sum_{i=1}^{n}Y_i-n\overline{Y}=0$ **が使われている。**

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

3.1.2 データについて

X_i，Y_i **は** i **番目のデータを表す。**

1. **タイム・シリーズ (時系列)・データ： 添え字** i **が時間を表す (**i **年，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)。**

3.2　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。

3.2.1　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。

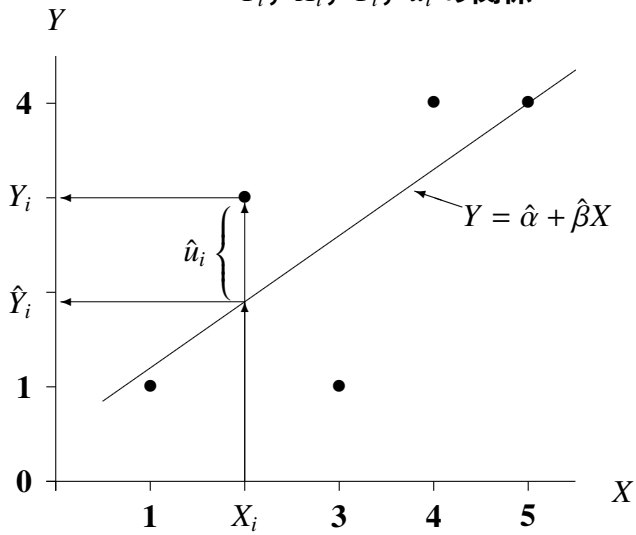
上の式は回帰モデル（または，回帰式）と呼ばれる。切片 α と傾き β をデータ $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ から求める（推定する）ことを考える。

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。

Y_i，X_i，$\hat{Y}_i$，$\hat{u}_i$ の関係



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

$\hat{Y}_i$ は Y_i の予測値と呼ばれる。

3.2.2　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，

$$\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0, \tag{3.1}$$

$$\sum_{i=1}^{n} X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0, \tag{3.2}$$

を満たす。

さらに，

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i \tag{3.3}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha} \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 \tag{3.4}$$

(3.3) 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，

$$\overline{Y} = \hat{\alpha} + \hat{\beta} \overline{X} \tag{3.5}$$

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 (3.5) 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，

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{S_{XY}}{S_X^2} \tag{3.6}$$

が得られ，$\hat{\alpha}$ は (3.5) 式から，

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \tag{3.7}$$

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。

前述の $\hat{\alpha}$，$\hat{\beta}$ の連立方程式：

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

を行列表示することによって，

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

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

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

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

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，**

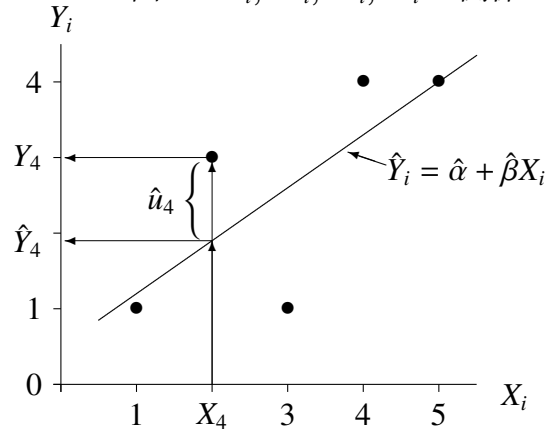
$$\hat{Y}_i = 0.5 + 0.7 X_i,$$

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。

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**			

Y_i，X_i，$\hat{Y}_i$，$\hat{u}_i$ の関係が図 3.1 に描かれている。

図 3.1: Y_i，X_i，$\hat{Y}_i$，$\hat{u}_i$ の関係



$\hat{Y}_i$ 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。

3.2.3　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，(3.1) 式，(3.2) 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**						