1.2. データの回帰分析

y　　目的変数
$\boldsymbol{x} = (x_1, \cdots, x_p)'$:　説明変数

データ：$(\boldsymbol{x}_i, y_i),\ \ i = 1, \cdots, n$
（標本）　$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})',\ \ (i = 1, \cdots, n)$

行列記号で、$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$, $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

統計量：
- $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$
- $\overline{y}$
- $S = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'$: 標本分散行列
- $s^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2$
- $\boldsymbol{s} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\overline{y}_i - \overline{y})$: 標本共分散ベクトル

y	x_1	$\cdots$	x_p	$\boldsymbol{x}'$
y_1	x_{11}	$\cdots$	x_{1p}	$\boldsymbol{x}_1'$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	x_{n1}	$\cdots$	x_{np}	$\boldsymbol{x}_n'$
$\overline{y}$	$\overline{x}_1$	$\cdots$	$\overline{x}_p$	$\overline{\boldsymbol{x}}'$

仮定：
- S : 正値
- $s > 0$

[注意] : $\text{rank } S \leqq \min(n-1, p)$
より、S : 正値の仮定は、$n \geqq p+1$ を要求する。

理由：$X_R := X - \mathbf{1_n}\overline{\boldsymbol{x}}' = \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} - \begin{bmatrix} \overline{\boldsymbol{x}}' \\ \vdots \\ \overline{\boldsymbol{x}}' \end{bmatrix} = \begin{bmatrix} (\boldsymbol{x}_1 - \overline{\boldsymbol{x}})' \\ \vdots \\ (\boldsymbol{x}_n - \overline{\boldsymbol{x}})' \end{bmatrix}$ とおくと、
$S = \frac{1}{n} X_R' X_R,\ p = \text{rank } S = \text{rank } X_R \leqq n-1$ より分かる。

Prop.1.2.1.　$Q(\alpha, \boldsymbol{\beta}) := \sum_{i=1}^{n} (y_i - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_i)^2$

データにおいて y を $\alpha + \boldsymbol{\beta}'\boldsymbol{x}$ で近似したときの誤差

問題
- $\min\limits_{\alpha, \boldsymbol{\beta}} Q(\alpha, \boldsymbol{\beta})$
- 条件　$\alpha \in \boldsymbol{R}^1,\ \boldsymbol{\beta} : p \times 1$

$\Longrightarrow$

解
- $\alpha = \widehat{\alpha} := \overline{y} - \widehat{\boldsymbol{\beta}}'\overline{\boldsymbol{x}}$
- $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} := S^{-1}\boldsymbol{s}$

最小値 $= n(s^2 - \boldsymbol{s}' S^{-1} \boldsymbol{s})$

証明：r.v. $\boldsymbol{X},\ Y$ を、次のように定義する。

$$Pr\{\boldsymbol{X}=\boldsymbol{x}_i,\ Y=y_i\}=\frac{1}{n}\qquad (i=1,\cdots,n)$$

このとき、$\boldsymbol{\mu}=E(\boldsymbol{X})=\dfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i=\overline{\boldsymbol{x}},\ \nu=E(Y)=\overline{y},$

$$\Sigma=\mathrm{Var}(\boldsymbol{X})=E\{\boldsymbol{X}-E(\boldsymbol{X})\}\{\boldsymbol{X}-E(\boldsymbol{X})\}'=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})'$$

$\boldsymbol{\sigma}=\mathrm{Cov}(\boldsymbol{X},Y)=\boldsymbol{s},\ \sigma^2=\mathrm{Var}(Y)=s^2$

より、

$$Q_0(\alpha,\boldsymbol{\beta})=E(Y-\alpha-\boldsymbol{\beta}'\boldsymbol{X})^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\alpha-\boldsymbol{\beta}'\boldsymbol{x}_i)^2=\frac{1}{n}Q(\alpha,\boldsymbol{\beta})$$

よって、Prop.1.1.1. より、

解　$\alpha=\alpha_0=\nu-\boldsymbol{\beta}'\boldsymbol{\mu}=\overline{y}-\boldsymbol{\beta_0}'\overline{\boldsymbol{x}}=\widehat{\alpha}$ とする。

$\boldsymbol{\beta}=\boldsymbol{\beta}_0=\Sigma^{-1}\boldsymbol{\sigma}=S^{-1}\boldsymbol{s}=\widehat{\boldsymbol{\beta}}$

Q の最小値は、$n\left(\sigma^2-\boldsymbol{\sigma}'\Sigma^{-1}\boldsymbol{\sigma}\right)=n\left(s^2-\boldsymbol{s}'S^{-1}\boldsymbol{s}\right)$　　**証明終**

Def.1.2.1.　$y=\widehat{\alpha}+\widehat{\boldsymbol{\beta}}_0'\boldsymbol{x}$　標本回帰関数（標本回帰平面）

$\widehat{y}_i:=\widehat{\alpha}+\widehat{\boldsymbol{\beta}}'\boldsymbol{x_i}$	回帰
$e_i:=y_i-\widehat{y}_i$	残差
$R:=\sqrt{\boldsymbol{s}'S^{-1}\boldsymbol{s}}/s$	標本重相関係数
R^2	寄与率（標本決定係数）

$0\leqq R\leqq 1$

$\left(\mathrm{Min}\ Q=\sigma^2(1-R^2)\ \geqq\ 0\right)$

Prop.1.2.2.　(i) $\dfrac{1}{n}\sum_{i=1}^{n}\widehat{y}_i=\overline{y}$：標本平均、　$\dfrac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2=s^2R^2$：標本分散

(ii) $\dfrac{1}{n}\sum_{i=1}^{n}e_i=0,\ \ \dfrac{1}{n}\sum_{i=1}^{n}e_i^2=s^2(1-R^2)$

(iii) $\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})e_i=0$：標本共分散

証明：(i) $\widehat{y}_i=\overline{y}+\widehat{\boldsymbol{\beta}}'(\boldsymbol{x_i}-\overline{\boldsymbol{x}})$

ゆえに、$\dfrac{1}{n}\sum_{i=1}^{n}\widehat{y}_i=\overline{y}$

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2=\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\boldsymbol{\beta}}'(\boldsymbol{x_i}-\overline{\boldsymbol{x}})\right\}^2$$
$$=\frac{1}{n}\widehat{\boldsymbol{\beta}}'\sum_{i=1}^{n}(\boldsymbol{x_i}-\overline{\boldsymbol{x}})(\boldsymbol{x_i}-\overline{\boldsymbol{x}})'\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}'S\widehat{\boldsymbol{\beta}}=\boldsymbol{s}'S^{-1}\boldsymbol{s}=s^2R^2$$

(ii) $e_i=(y_i-\overline{y})-\widehat{\boldsymbol{\beta}}'(\boldsymbol{x_i}-\overline{\boldsymbol{x}})$

ゆえに、$\frac{1}{n}\sum_{i=1}^{n}e_i=0$

$\frac{1}{n}\sum_{i=1}^{n}e_i^2=Q(\widehat{\alpha},\widehat{\boldsymbol{\beta}})=s^2(1-R^2)$

(iii) $\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})e_i=\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}'(\boldsymbol{x_i}-\overline{\boldsymbol{x}})\left[(y_i-\overline{y})-\widehat{\boldsymbol{\beta}}'(\boldsymbol{x_i}-\overline{\boldsymbol{x}})\right]=\widehat{\boldsymbol{\beta}}'n(\boldsymbol{s}-S\widehat{\boldsymbol{\beta}})=0$　　証明終

[注意]　要因の分解

変量	$y_i=$	$\widehat{y}_i$	$+$	e_i
標本平均	$\overline{y}=$	$\overline{y}$	$+$	0
標本分散	$s^2=$	s^2R^2	$+$	$s^2(1-R^2)$
		回帰		残差

Prop.1.2.3.　$\boldsymbol{\beta}'\boldsymbol{x_i}$ $(i=1,\cdots,n)$ と y_i $(i=1,\cdots,n)$ の標本共分散を $\mathrm{Cov}(\boldsymbol{\beta}'\boldsymbol{x},y)$ で表すことにする。$\mathrm{Var}(\boldsymbol{\beta}'\boldsymbol{x})$, $\mathrm{Var}(y)$ も同様。

$$\mathrm{cor}(\boldsymbol{\beta}'\boldsymbol{x},y)=\mathrm{Cov}(\boldsymbol{\beta}'\boldsymbol{x},y)/\left(\sqrt{\mathrm{Var}(\boldsymbol{\beta}'\boldsymbol{x})}\sqrt{\mathrm{Var}(y)}\right).$$

$$\text{問題}\begin{cases}\max\limits_{\boldsymbol{\beta}}\ \mathrm{cor}(\boldsymbol{\beta}'\boldsymbol{x},y)\\ \text{条件}\quad \boldsymbol{\beta}:p\times 1\end{cases}$$

$\Longrightarrow$

解：$\boldsymbol{\beta}\propto\widehat{\boldsymbol{\beta}}$

最大値 $=R$

理由：Prop.1.2.1 の $\boldsymbol{X}$, Y を使うと、

$$\mathrm{Var}(\boldsymbol{\beta}'\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}'\boldsymbol{x_i}-\boldsymbol{\beta}'\overline{\boldsymbol{x}})^2=\mathrm{Var}(\boldsymbol{\beta}'\boldsymbol{X})$$

$$\mathrm{Var}(y)=\mathrm{Var}(Y)$$

$$\mathrm{Cov}(\boldsymbol{\beta}'\boldsymbol{x},y)=\mathrm{Cov}(\boldsymbol{\beta}'\boldsymbol{X},Y)$$

ゆえに、$\mathrm{cor}(\boldsymbol{\beta}'\boldsymbol{x},y)=\mathrm{cor}(\boldsymbol{\beta}'\boldsymbol{X},Y)$

よって、Prop.1.1.3 より、$\boldsymbol{\beta}\propto\boldsymbol{\beta}_0=\Sigma^{-1}\boldsymbol{\sigma}=S^{-1}\boldsymbol{s}=\widehat{\boldsymbol{\beta}}$　　理由終

2021.12.21(火 2)

ZOOM

これまでは、α と $\boldsymbol{\beta}$ の扱いが非対称であった。

$$y_i \quad \leftarrow \quad \alpha + \boldsymbol{\beta}'\boldsymbol{x_i}$$

（α：定数項、$\boldsymbol{\beta}$：偏回帰係数）

これを、$y_i \quad \leftarrow \quad \alpha \cdot 1 + \boldsymbol{\beta}'\boldsymbol{x_i}$

と考え、$\begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} 1 \\ \boldsymbol{x_i} \end{bmatrix}$ を改めて、$\boldsymbol{\beta}$, $\boldsymbol{x_i}$ と書けば、$y_i \quad \leftarrow \quad \boldsymbol{\beta}'\boldsymbol{x_i}$ となる。

逆に新しい記号において、$\boldsymbol{x_i}$ の第 1 成分が恒等的に 1 ならば、前の場合になる。その場合を、「**定数項のある場合**」という。

データ：$(\boldsymbol{x_i}, y_i), \ (i = 1, \cdots, n)$

行列記号で、X, $\boldsymbol{y}$

仮定：rank $X = p$　　このとき、X は **full rank** という。

[**注意**]：定数項がある場合、これを除いた $(p-1)$ 個の変数についての標本分散行列を S とすると、

$$full\ rank \quad \Longleftrightarrow \quad S：正値$$

Prop.1.2.4.　$Q(\boldsymbol{\beta}) := ||\boldsymbol{y} - X\boldsymbol{\beta}||^2$ とする。このとき、以下の性質が成り立つ。

(i) $Q(\boldsymbol{\beta}) = Q(\widehat{\boldsymbol{\beta}}) + ||X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})||^2$

ここで、$\widehat{\boldsymbol{\beta}} := (X'X)^{-1}X'\boldsymbol{y}$

(ii) 問題 $\begin{cases} \min\limits_{\boldsymbol{\beta}} \ Q(\boldsymbol{\beta}) \\ 条件 \quad \boldsymbol{\beta} : p \times 1 \end{cases}$

$\Longrightarrow$

解：$\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$

最小値 $= \boldsymbol{y}'(I - \Pi_X)\boldsymbol{y}$

さらに、$\Pi_X := X(X'X)^{-1}X'$: $n \times n$, 対称、巾等、rank $\Pi_X = p$

証明：(i) $Q(\boldsymbol{\beta}) = ||(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) - X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})||^2$

$= ||\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}||^2 + ||X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})||^2$

なぜならば、積和 $= 2\{X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\}'(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = 2(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\underline{X'(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})} = 0$

（下線部は、$\mathbf{0}$）

(ii) $Q(\boldsymbol{\beta})$ を最小にする $\boldsymbol{\beta}$ は、$||X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})||^2 = 0$ の解

$(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'X'X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) = 0$

ここで、$X'X$ は正値であるから、$\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ に限る。

$\boldsymbol{e} := \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$

$= \boldsymbol{y} - X(X'X)^{-1}X'\boldsymbol{y} = \{I_n - X(X'X)^{-1}X'\}\boldsymbol{y} = (I - \Pi_X)\boldsymbol{y}$