

回帰分析

データ数を n 、説明変数の数を m とする。

各データ ($i=1\sim n$) について $y_i = a_0 + \sum_{j=1}^m a_j x_{i,j} + \varepsilon_i$ とする。

目的変数の標本値ベクトルを $\{y\} = \begin{Bmatrix} y_1 \\ \vdots \\ y_n \end{Bmatrix}$ 、

説明変数の標本値を $[X] = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,m} \end{bmatrix}$

未知のパラメータを $\{a\} = \begin{Bmatrix} a_0 \\ \vdots \\ a_m \end{Bmatrix}$ 、

誤差を $\{\varepsilon\} = \begin{Bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{Bmatrix}$ として、任意の i, j ($i, j = 1\sim n$) について

- $E(\varepsilon_i) = 0$ 誤差の平均は 0
- $E(\varepsilon_i^2) = \sigma^2$ 誤差の分散は同じ
- $E(\varepsilon_i \varepsilon_j) = 0$ ($i \neq j$) 誤差 ε_i と ε_j とは独立
- ε_i は正規分布をする

とする。

上式は $\{y\} = [X] \{a\} + \{\varepsilon\}$ となる。

ここで、 $Q = \{\varepsilon\}^t \{\varepsilon\} = (\{y\} - [X] \{a\})^t (\{y\} - [X] \{a\})$

が 最少となる $\{a\}$ を求める。

$$\begin{aligned} Q &= (\{y\} - [X] \{a\})^t (\{y\} - [X] \{a\}) \\ &= (\{y\}^t - \{a\}^t [X]^t) (\{y\} - [X] \{a\}) \\ &= \{y\}^t \{y\} - 2\{a\}^t [X]^t \{y\} + \{a\}^t [X]^t [X] \{a\} \end{aligned}$$

$$\text{となり } \frac{\partial Q}{\partial \{a\}} = -2[X]^t \{y\} + 2[X]^t [X] \{a\} = \{0\}$$

として、結局、 $[X]^t [X] \{a\} = [X]^t \{y\}$ となる。

つまり、 $\{a\} = ([X]^t [X])^{-1} [X]^t \{y\}$ である。

上記で計算された $\{a\}$ を用いた 推定値を

$$Y_i = a_0 + \sum_{j=1}^m a_j x_{i,j} \quad (i = 1 \sim n)$$

また、 $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ 、 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ として、

$$\text{重相関係数 } R = \frac{\sum_{i=1}^n (y_i - \bar{y})(Y_i - \bar{Y})}{\sqrt{(\sum_{i=1}^n (y_i - \bar{y})^2) (\sum_{i=1}^n (Y_i - \bar{Y})^2)}}$$

決定係数は R^2

自由度調整済み決定係数は $1.0 - \frac{n-1}{n-m-1} (1 - R^2)$ である

推定の有効性を分散分析を利用して検討する。

目的変数の全変動 $\sum_{i=1}^n (y_i - \bar{y})^2$ は、以下のように表現できる

$$\begin{aligned} (A) = \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - Y_i + Y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - Y_i + Y_i - \bar{Y})^2 \quad (\because \bar{y} = \bar{Y}) \\ &= \underbrace{\sum_{i=1}^n (y_i - Y_i)^2}_{(B)} + \underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{(C)} \quad (\because \sum_{i=1}^n (y_i - Y_i)(Y_i - \bar{Y}) = 0) \end{aligned}$$

(A) は一定で、(B) は残差の合計 $(= \sum_{i=1}^n \varepsilon^2)$ である。

つまり、(B) が小さければ 推定精度が良いことになる。

(B) を自由度 $n - m - 1$ で割った値を $E = (B) / (n - m - 1)$

(C) を自由度 m で割った値を $R = (C) / m$ とすると、

$F = R / E$ は 自由度 $(m, n - m - 1)$ の F 分布に従うので、

この F 値に基づく p 値が有意水準以下であれば、推定の有意性が認められる。

つまり帰無仮説 : $a_j = 0$ ($j = 1 \sim m$) が棄却される。

また 標準誤差は \sqrt{E} である。

次に偏回帰係数の有効性について検討する。

説明変数 x_i, x_j との偏差積和を

$$a_{i,j} = \sum_{k=1}^n (x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j) \quad (\bar{x}_i = \frac{1}{n} \sum_{k=1}^n x_{k,i})$$

として 偏差積和行列

$$[A] = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,m} \end{bmatrix} \quad a_{i,j} = a_{j,i}$$

$$[A]^{-1} = \begin{bmatrix} a^{1,1} & a^{1,2} & \cdots & a^{1,m} \\ a^{2,1} & a^{2,2} & \cdots & a^{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ a^{m,1} & a^{m,2} & \cdots & a^{m,m} \end{bmatrix}$$

$i = 1 \sim m$ については

$$t_i = \frac{a_i - 0.0}{\sqrt{\frac{a^{i,i} E}{n}}} \text{ が自由度 } n - m - 1 \text{ の } t \text{ 分布に従うので、}$$

計算された t 値から有意確率 p を計算して、有意かどうかを検定できる。

$i = 0$ については

$$t_0 = \frac{a_0 - 0.0}{\sqrt{\frac{1}{n} \left(1 + \sum_{i=1}^m \sum_{j=1}^m \bar{x}_i \bar{x}_j a^{i,j} \right) E}}$$

が自由度 $n - m - 1$ の t 分布に従うので、

計算された t 値から有意確率 p を計算して、有意かどうかを検定できる。

また、 $r_{i,j}$ を x_i と x_j との相関係数、 $r_{i,y}$ を x_i と y との相関係数として、

$$[R] = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} & r_{1,y} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} & r_{2,y} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m,1} & r_{m,2} & \cdots & r_{m,m} & r_{m,y} \\ r_{y,1} & r_{y,2} & \cdots & r_{y,m} & r_{y,y} \end{bmatrix}$$

$$[R]^{-1} = \begin{bmatrix} r^{1,1} & r^{1,2} & \cdots & r^{1,m} & r^{1,y} \\ r^{2,1} & r^{2,2} & \cdots & r^{2,m} & r^{2,y} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r^{m,1} & r^{m,2} & \cdots & r^{m,m} & r^{m,y} \\ r^{y,1} & r^{y,2} & \cdots & r^{y,m} & r^{y,y} \end{bmatrix}$$

としたとき、変数 x_i と y との偏相関係数 $r_{iy \cdot 123 \dots i-1, i+1 \dots m}$ は以下のように計算する。

$$r_{iy \cdot 123 \dots i-1, i+1 \dots m} = \frac{-r^{i,y}}{\sqrt{r^{i,i}} \sqrt{r^{y,y}}}$$