応用システム工学
第三回　回帰分析

平成23年04月22日

線形重回帰

目的変数yの，説明変数$x_1,x_2,\cdots,x_p$に対する線形重回帰式

$$y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_p x_p$$

重回帰係数　　$\hat{a}_j \quad j = 1,2,\cdots,p$

重回帰式を分散共分散行列で表す

$$s_{jl} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ji} - \bar{x}_j\right)\left(x_{li} - \bar{x}_l\right)$$

$$j,l = 1,2,\cdots,p$$

$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & & & \\ s_{j1} & s_{j2} & & s_{jp} \\ \vdots & & & \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} s_{y1} \\ \vdots \\ s_{yj} \\ \vdots \\ s_{yp} \end{bmatrix}$$

分散共分散行列V　　yのxに対する共分散

$$\hat{a}_0 = \bar{y} - \left(\hat{a}_1\bar{x}_1 + \hat{a}_2\bar{x}_2 + \cdots + \hat{a}_p\bar{x}_p\right)$$

で　　を求めればよい

重回帰式の予測誤差の標準偏差

- 予測誤差の標準偏差

$$s_e = \sqrt{\frac{1}{n-(p+1)}\sum_{i=1}^{n}(e_i-\bar{e})^2} = \sqrt{\frac{1}{n-(p+1)}\sum_{i=1}^{n}e_i^{\,2}}$$

  - ただし $\bar{e} = \frac{1}{n}\sum_{i=1}^{n}e_i = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \left(\hat{a}_0 + \hat{a}_1x_{1i} + \hat{a}_2x_{2i} + \cdots + \hat{a}_px_{pi}\right)\right\}$

$$\hat{a}_0 = \bar{y} - \left(\hat{a}_1\bar{x}_1 + \hat{a}_2\bar{x}_2 + \cdots + \hat{a}_p\bar{x}_p\right) \quad \text{より}$$

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \left[\bar{y} - \left(\hat{a}_1\bar{x}_1 + \hat{a}_2\bar{x}_2 + \cdots + \hat{a}_p\bar{x}_p\right) + \hat{a}_1x_{1i} + \hat{a}_2x_{2i} + \cdots + \hat{a}_px_{pi}\right]\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{(y_i-\bar{y}) - \hat{a}_1(x_{1i}-\bar{x}_1)\cdots\hat{a}_p\left(x_{pi}-\bar{x}_p\right)\right\} = 0$$

$\frac{\partial}{\partial a_0}F(a_0,a_1,\cdots,a_p) = 0$ に一致

重相関係数

- 重回帰式による予測値

$$Y_i = \hat{a}_0 + \hat{a}_1\bar{x}_1 + \hat{a}_2\bar{x}_2 + \cdots + \hat{a}_p\bar{x}_p$$

- 予測値Yiは説明変数で表される
- 目的変数yと予測値Yの単相関係数 r_{yY}
  →目的変数yと説明変数$x_1,x_2,\cdots,x_p$の重相関係数

$$r_{y\cdot 12\cdots p} = \frac{s_{yY}}{\sqrt{s_{yy}s_{YY}}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

  - 回帰平面(p+1)次元に近いかどうかを表す

重相関係数

- 重相関係数のとる範囲は？

$$r_{y\cdot 12\cdots p}=\frac{s_{yY}}{\sqrt{s_{yy}s_{YY}}}=\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

– 分子について考える

$$s_{yY}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(Y_i-\bar{Y})$$

重相関係数

– 目的変数と予測値の平均

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \qquad \bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{1}{n}\sum_{i=1}^{n}(y_i-e_i)=\frac{1}{n}\sum_{i=1}^{n}y_i=\bar{y}$$

– 分子

$$\begin{aligned}s_{yY}&=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(Y_i-\bar{Y})=\frac{1}{n}\sum_{i=1}^{n}(Y_i+e_i-\bar{Y})(Y_i-\bar{Y})\\&=\frac{1}{n}\sum_{i=1}^{n}\left\{(Y_i-\bar{Y})^2+e_i(Y_i-\bar{Y})\right\}\\&=\frac{1}{n}\sum_{i=1}^{n}\left\{(Y_i-\bar{Y})^2+e_i(a_0+a_1x_{1i}+\cdots+a_px_{pi}-\bar{Y})\right\}\\&=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2=s_{YY}\geq 0\end{aligned} \Rightarrow r_{y\cdot 12\cdots p}=\frac{s_{yY}}{\sqrt{s_{yy}s_{YY}}}\geq 0$$

重相関係数

– 回帰式の残差平方和を最小にする係数の条件

$$\frac{\partial F}{\partial a_0} = \frac{\partial}{\partial a_0}\sum_{i=1}^{n}\left\{y_i - \left(a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi}\right)\right\}^2$$

$$= -2\sum_{i=1}^{n}\left\{y_i - \left(a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi}\right)\right\} = \sum_{i=1}^{n} e_i = 0$$

$$\frac{\partial F}{\partial a_j} = \frac{\partial}{\partial a_j}\sum_{i=1}^{n}\left\{y_i - \left(a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi}\right)\right\}^2$$

$$= -2\sum_{i=1}^{n} x_{ji}\left\{y_i - \left(a_0 + a_1 x_{1i} + \cdots + a_p x_{pi}\right)\right\} = -2\sum_{i=1}^{n} x_{ji} e_i = 0$$

重相関係数

- 両辺二乗 $r_{y\cdot 12\cdots p}{}^2 = \dfrac{s_{yY}{}^2}{s_{yy}s_{YY}} = \dfrac{s_{YY}{}^2}{s_{yy}s_{YY}} = \dfrac{s_{YY}}{s_{yy}}$
- 目的変数･予測値の分散の関係を求める

$$s_{yy} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i + e_i - \bar{Y}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(Y_i - \bar{Y}\right)^2 + 2e_i\left(Y_i - \bar{Y}\right) + e_i{}^2\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(Y_i - \bar{Y}\right)^2 + 2e_i\left(a_0 + a_1 x_{1i} + \cdots + a_p x_{pi} - \bar{Y}\right) + e_i{}^2\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(Y_i - \bar{Y}\right)^2 + e_i{}^2\right\} = s_{YY} + \frac{1}{n}\sum_{i=1}^{n} e_i{}^2$$

重相関係数

$$\frac{1}{n}\sum_{i=1}^{n} e_i^{\,2} = s_{yy} - s_{YY} = s_{yy}\left(1 - \frac{s_{YY}}{s_{yy}}\right)$$

$$= s_{yy}\left(1 - r_{y\cdot 12\cdots p}^{\,2}\right) \geq 0$$

$$1 - r_{y\cdot 12\cdots p}^{\,2} \geq 0$$

$$-1 \leq r_{y\cdot 12\cdots p} \leq 1$$

$$r_{y\cdot 12\cdots p} = \frac{s_{yY}}{\sqrt{s_{yy}s_{YY}}} \geq 0$$ より

$$0 \leq r_{y\cdot 12\cdots p} \leq 1$$

偏相関係数

- 目的変数y,x_1を説明変数$x_2,x_3,\cdots,x_p$から予測する二つの重回帰モデル

$$\begin{cases} y_i = c_0 + c_2 x_{2i} + \cdots + c_p x_{pi} + e_i \\ x_{1i} = d_0 + d_2 x_{2i} + \cdots + d_p x_{pi} + e_i' \end{cases}$$

  - 予測誤差の平方和

$$\begin{cases} F(c_0, c_2, \cdots, c_p) = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left\{ y_i - \left(c_0 + c_2 x_{2i} + \cdots + c_p x_{pi}\right)\right\}^2 \\ F'(d_0, d_2, \cdots, d_p) = \sum_{i=1}^{n} e_i'^{\,2} = \sum_{i=1}^{n} \left\{ x_{1i} - \left(d_0 + d_2 x_{2i} + \cdots + d_p x_{pi}\right)\right\}^2 \end{cases}$$

偏相関係数

– 予測誤差の平方和の最小化(最小二乗法)

$$\begin{cases} \dfrac{\partial}{\partial c_0} F(c_0, c_2, \cdots, c_p) = 0 \\ \dfrac{\partial}{\partial c_2} F(c_0, c_2, \cdots, c_p) = 0 \\ \vdots \\ \dfrac{\partial}{\partial c_p} F(c_0, c_2, \cdots, c_p) = 0 \end{cases} \qquad \begin{cases} \dfrac{\partial}{\partial d_0} F'(d_0, d_2, \cdots, d_p) = 0 \\ \dfrac{\partial}{\partial d_2} F'(d_0, d_2, \cdots, d_p) = 0 \\ \vdots \\ \dfrac{\partial}{\partial d_p} F'(d_0, d_2, \cdots, d_p) = 0 \end{cases}$$

偏相関係数

– 予測誤差の平方和の最小化(最小二乗法)

$$\begin{cases} -2\sum_{i=1}^{n} \{y_i - (c_0 + c_2 x_{2i} + \cdots + c_p x_{pi})\} = 0 \\ -2\sum_{i=1}^{n} x_{2i} \{y_i - (c_0 + c_2 x_{2i} + \cdots + c_p x_{pi})\} = 0 \\ \vdots \\ -2\sum_{i=1}^{n} x_{pi} \{y_i - (c_0 + c_2 x_{2i} + \cdots + c_p x_{pi})\} = 0 \end{cases}$$

偏相関係数

– 予測誤差の平方和の最小化(最小二乗法)

$$\begin{cases} -2\sum_{i=1}^{n}\left\{x_{1i}-\left(d_0+d_2x_{2i}+\cdots+d_px_{pi}\right)\right\}=0 \\ -2\sum_{i=1}^{n}x_{2i}\left\{x_{1i}-\left(d_0+d_2x_{2i}+\cdots+d_px_{pi}\right)\right\}=0 \\ \vdots \\ -2\sum_{i=1}^{n}x_{pi}\left\{x_{1i}-\left(d_0+d_2x_{2i}+\cdots+d_px_{pi}\right)\right\}=0 \end{cases}$$

偏相関係数

- $x_2,\cdots,x_p$の分散共分散行列

$$V=\begin{bmatrix} s_{22} & s_{23} & \cdots & s_{2l} & \cdots & s_{2p} \\ s_{32} & s_{33} & & & & s_{3p} \\ \vdots & & & & & \vdots \\ s_{j2} & s_{j3} & \cdots & s_{jl} & \cdots & s_{jp} \\ \vdots & & & & & \vdots \\ s_{p2} & s_{p3} & \cdots & s_{pl} & \cdots & s_{pp} \end{bmatrix} \qquad s_{jl}=\frac{1}{n}\sum_{i=1}^{n}\left(x_{ji}-\bar{x}_j\right)\left(x_{li}-\bar{x}_l\right)$$

$$j,l=1,2,\cdots,p$$

- y, x_1と$x_2,\cdots,x_p$の共分散

$$s_{yj}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(x_{ji}-\bar{x}_j\right) \quad s_{lj}=\frac{1}{n}\sum_{i=1}^{n}\left(x_{li}-\bar{x}_l\right)\left(x_{ji}-\bar{x}_j\right)$$

$$j=2,3,\cdots,p$$

偏相関係数

$$\begin{bmatrix} s_{22} & s_{23} & \cdots & s_{2p} \\ \vdots & & & \\ s_{j2} & s_{j3} & & s_{jp} \\ \vdots & & & \\ s_{p2} & s_{p3} & \cdots & s_{pp} \end{bmatrix} \begin{bmatrix} \hat{c}_2 \\ \hat{c}_3 \\ \vdots \\ \hat{c}_p \end{bmatrix} = \begin{bmatrix} s_{y2} \\ \vdots \\ s_{yj} \\ \vdots \\ s_{yp} \end{bmatrix} \qquad \begin{bmatrix} s_{22} & s_{23} & \cdots & s_{2p} \\ \vdots & & & \\ s_{j2} & s_{j3} & & s_{jp} \\ \vdots & & & \\ s_{p2} & s_{p3} & \cdots & s_{pp} \end{bmatrix} \begin{bmatrix} \hat{d}_2 \\ \hat{d}_3 \\ \vdots \\ \hat{d}_p \end{bmatrix} = \begin{bmatrix} s_{12} \\ \vdots \\ s_{1j} \\ \vdots \\ s_{1p} \end{bmatrix}$$

$$\begin{cases} \hat{c}_0 = \bar{y} - \left(\hat{c}_2 \bar{x}_2 + \hat{c}_3 \bar{x}_3 + \cdots + \hat{c}_p \bar{x}_p \right) \\ \hat{d}_0 = \bar{x}_1 - \left(\hat{d}_2 \bar{x}_2 + \hat{d}_3 \bar{x}_3 + \cdots + \hat{d}_p \bar{x}_p \right) \end{cases}$$

目的変数y, x_1の，説明変数$x_2, x_3, \cdots, x_p$に対する線形重回帰式

$$\begin{cases} y = \hat{c}_0 + \hat{c}_2 x_2 + \hat{c}_3 x_3 + \cdots + \hat{c}_p x_p \\ x_1 = \hat{d}_0 + \hat{d}_2 x_2 + \hat{d}_3 x_3 + \cdots + \hat{d}_p x_p \end{cases}$$

重回帰係数 $\hat{c}_j, \hat{d}_j \qquad j = 2,3,\cdots, p$

偏相関係数

- 予測誤差

$$\begin{cases} u_i = y_i - \left(c_0 + c_2 x_{2i} + \cdots + c_p x_{pi} \right) \\ v_i = x_{1i} - \left(d_0 + d_2 x_{2i} + \cdots + d_p x_{pi} \right) \end{cases} \qquad i = 1,2,\cdots, n$$

  - 予測誤差u,vの単相関係数

$$r_{y1\bullet 23\cdots p} = \frac{s_{uv}}{\sqrt{s_{uu} s_{vv}}} \qquad s_{uu} = \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})^2 \quad s_{vv} = \frac{1}{n} \sum_{i=1}^{n} (v_i - \bar{v})^2$$

$$s_{uv} = \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})(v_i - \bar{v}) \qquad \bar{u} = \frac{1}{n} \sum_{i=1}^{n} u_i \qquad \bar{v} = \frac{1}{n} \sum_{i=1}^{n} v_i$$

y, x_1から，$x_2, x_3, \cdots, x_p$の回帰が消去された時の偏相関係数
→($x_2, x_3, \cdots, x_p$)の影響を除いたy, x_1の相関係数

偏相関係数

- 偏相関の意味
  - yの残差uはx2,・・・,xpに依存する変動をyから除いたもの
  - X1の残差vはx2,・・・,xpに依存する変動をx1から除いたもの
  - 予測誤差u,vの相関係数は，yとx1から各々x2,・・・,xpに依存する変動を除いた相関係数

標本と母数の関係
単回帰

- 母回帰係数 a_0, a_1
  - 予測値 $a_0 + a_1 x_i$
  - 標本 y_i, x_i
  - 誤差 $e_i = y_i - (a_0 + a_1 x_i)$
    (残差ではない)
- 標本回帰係数 $\hat{a}_0, \hat{a}_1$
  - 残差 $y_i - (\hat{a}_0 + \hat{a}_1 x_i)$
- 母集団を求めることは難しいので，標本から母集団を推定する

母回帰係数の推定

- 仮定
  - 誤差の期待値　$E[e_i]=0$
  - 誤差の分散　$V[e_i]=\sigma^2$
  - 誤差は無相関　$Cov[e_i,e_j]=0(i\neq j)$
  - 誤差は正規分布　$N(0,\sigma^2)$
- 標本回帰係数の期待値との関係を求める
  - 標本回帰係数　$\hat{a}_0,\hat{a}_1$

母回帰係数の推定

- 標本回帰係数　$\hat{a}_1$ の期待値 $E[\hat{a}_1]$

$$\hat{a}_1=\frac{s_{xy}}{s_{xx}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$y_i=a_0+a_1x_i+e_i$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{n}(a_0+a_1x_i+e_i)$$

$$=a_0+a_1\bar{x}$$

$$=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(a_0+a_1x_i+e_i-[a_0+a_1\bar{x}])}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(a_1[x_i-\bar{x}]+e_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=a_1+\frac{\sum_{i=1}^{n}(x_i-\bar{x})e_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ の期待値$E[\hat{a}_1]$

$$E[\hat{a}_1] = E\left[a_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] = E[a_1] + E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

  – 確率変数 e_i なので，第二項は0

$$E[\hat{a}_1] = E[a_1] = a_1$$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ の分散 $V[\hat{a}_1]$

$$V[\hat{a}_1] = E\left[(\hat{a}_1 - E[\hat{a}_1])^2\right] = E\left[(\hat{a}_1 - a_1)^2\right]$$

$$= E\left[\left(a_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - a_1 \right)^2 \right]$$

$$= E\left[\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)^2 \right]$$

母回帰係数の推定

- つづき

$$V[\hat{a}_1] = E\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\bar{x})(x_j-\bar{x})e_i e_j}{\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^2}\right]$$

$$Cov[e_i, e_j]$$
$$= E[(e_i - E[e_i])(e_j - E[e_j])]$$
$$= E[e_i e_j] = 0$$

$$= E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 e_i^{\,2}}{\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^2}\right]$$

$$V[e_i] = E[e_i^{\,2}] - E[e_i]^2$$
$$= E[e_i^{\,2}] = \sigma^2$$

$$= \frac{\sigma^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$