応用システム工学
第六回　回帰分析

平成24年07月06日
重回帰分析

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

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

偏相関係数

– 予測誤差の平方和の最小化(最小二乗法)

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

偏相関係数

– 予測誤差の平方和の最小化(最小二乗法)
- y_iに対して

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

偏相関係数

– 予測誤差の平方和の最小化(最小二乗法)
  - x_1に対して

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

偏相関係数

- $x_2,\cdots,x_p$の分散共分散行列

$$V=\begin{bmatrix} s_{22} & s_{23} & \cdots & s_{2l} & \cdots & s_{2p} \\ s_{32} & s_{33} & & & & s_{3p} \\ \vdots & & & & & \vdots \\ s_{j2} & s_{j3} & \cdots & s_{jl} & \cdots & s_{jp} \\ \vdots & & & & & \vdots \\ s_{p2} & s_{p3} & \cdots & s_{pl} & \cdots & s_{pp} \end{bmatrix}$$

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

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

- 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,･･･,x_pに対する線形重回帰式

$$
\begin{cases} y = \hat{c}_0 + \hat{c}_2x_2 + \hat{c}_3x_3 + \cdots + \hat{c}_px_p \\ x_1 = \hat{d}_0 + \hat{d}_2x_2 + \hat{d}_3x_3 + \cdots + \hat{d}_px_p \end{cases}
$$

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

$$
\begin{cases} u_i = y_i - \left(c_0 + c_2x_{2i} + \cdots + c_px_{pi}\right) \\ v_i = x_{1i} - \left(d_0 + d_2x_{2i} + \cdots + d_px_{pi}\right) \end{cases} \qquad i = 1,2,\cdots,n
$$

  - 予測誤差u,vの単相関係数

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

$$
s_{uv} = \frac{1}{n}\sum_{i=1}^{n}\left(u_i - \bar{u}\right)\left(v_i - \bar{v}\right) \qquad \text{ただし} \quad \bar{u} = \frac{1}{n}\sum_{i=1}^{n}u_i \quad \bar{v} = \frac{1}{n}\sum_{i=1}^{n}v_i
$$

y, x_1から，x_2,x_3,･･･,x_pの回帰が消去された時の偏相関係数
→(x_2,x_3,･･･,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x_2,・・・,x_p依存する変動をyから除いたもの
  - X_1の残差vはx_2,・・・,x_pに依存する変動をx_1から除いたもの
  - 予測誤差u,vの相関係数は，yとx_1から各々x_2,・・・,x_p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

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

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

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

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

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

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ と母回帰係数 a_1 の関係

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

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

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

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

$$V[\hat{a}_1]=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ は誤差 e_i の一次関数として表される

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

→標本の誤差が正規分布なので標本回帰係数も正規分布する

$$\hat{a}_1 \sim N\left(a_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

  - 標準正規分布化 $\dfrac{\hat{a}_1 - a_1}{\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim N(0,1)$
    - 期待値を引いて分散で割る

→ただし，σ^2，a_1は未知数

大数の法則

- 標本数 → 用いる母集団の数
- 標本サイズ→ 母集団からとりだす標本の数
- 大数の法則→ 標本サイズを大きくすると，その期待値が母集団の期待値から離れた値となる確率は小さくなる
  - 標本平均の期待値 $E[\bar{X}_n]$ と分散 $V[\bar{X}_n]$ で評価
  - 標本$X_1,X_2,\cdots,X_n$は母集団と同じ確率分布（平均μ，分散σ^2）となる確率変数
  - 標本$X_1,X_2,\cdots,X_n$は独立